THE UNIVERSITY OF WARWICK

Library Declaration and Deposit Agreement

1. STUDENT DETAILS

Please complete the following:

Full name: Giorgos Minas

University ID number: 0852199

2. THESIS DEPOSIT

2.1 I understand that under my registration at the University, I am required to deposit my thesis with the University in BOTH hard copy and in digital format. The digital version should normally be saved as a single pdf file.

2.2 The hard copy will be housed in the University Library. The digital version will be deposited in the University's Institutional Repository (WRAP). Unless otherwise indicated (see 2.3 below) this will be made openly accessible on the Internet and will be supplied to the British Library to be made available online via its Electronic Theses Online Service (EThOS) service.
[At present, theses submitted for a Master's degree by Research (MA, MSc, LLM, MS or MMedSci) are not being deposited in WRAP and not being made available via EthOS. This may change in future.]

2.3 In exceptional circumstances, the Chair of the Board of Graduate Studies may grant permission for an embargo to be placed on public access to the hard copy thesis for a limited period. It is also possible to apply separately for an embargo on the digital version. (Further information is available in the *Guide to Examinations for Higher Degrees by Research*.)

2.4 *If you are depositing a thesis for a Master's degree by Research, please complete section (a) below. For all other research degrees, please complete both sections (a) and (b) below:*

(a) Hard Copy

I hereby deposit a hard copy of my thesis in the University Library to be made publicly available to readers (please delete as appropriate) EITHER immediately OR after an embargo period of 5 months/years as agreed by the Chair of the Board of Graduate Studies.

I agree that my thesis may be photocopied. YES / NO (*Please delete as appropriate*)

(b) Digital Copy

I hereby deposit a digital copy of my thesis to be held in WRAP and made available via EThOS.

Please choose one of the following options:

EITHER My thesis can be made publicly available online. YES / NO (*Please delete as appropriate*)

OR My thesis can be made publicly available only after 05/08/2018 [date] (Please give date)
YES / NO (*Please delete as appropriate*)

OR My full thesis cannot be made publicly available online but I am submitting a separately identified additional, abridged version that can be made available online.
YES / NO (*Please delete as appropriate*)

OR My thesis cannot be made publicly available online. YES / NO (*Please delete as appropriate*)

3. GRANTING OF NON-EXCLUSIVE RIGHTS

Whether I deposit my Work personally or through an assistant or other agent, I agree to the following:

Rights granted to the University of Warwick and the British Library and the user of the thesis through this agreement are non-exclusive. I retain all rights in the thesis in its present version or future versions. I agree that the institutional repository administrators and the British Library or their agents may, without changing content, digitise and migrate the thesis to any medium or format for the purpose of future preservation and accessibility.

4. DECLARATIONS

(a) I DECLARE THAT:

- I am the author and owner of the copyright in the thesis and/or I have the authority of the authors and owners of the copyright in the thesis to make this agreement. Reproduction of any part of this thesis for teaching or in academic or other forms of publication is subject to the normal limitations on the use of copyrighted materials and to the proper and full acknowledgement of its source.

- The digital version of the thesis I am supplying is the same version as the final, hard-bound copy submitted in completion of my degree, once any minor corrections have been completed.

- I have exercised reasonable care to ensure that the thesis is original, and does not to the best of my knowledge break any UK law or other Intellectual Property Right, or contain any confidential material.

- I understand that, through the medium of the Internet, files will be available to automated agents, and may be searched and copied by, for example, text mining and plagiarism detection software.

(b) IF I HAVE AGREED (in Section 2 above) TO MAKE MY THESIS PUBLICLY AVAILABLE DIGITALLY, I ALSO DECLARE THAT:

- I grant the University of Warwick and the British Library a licence to make available on the Internet the thesis in digitised format through the Institutional Repository and through the British Library via the EThOS service.

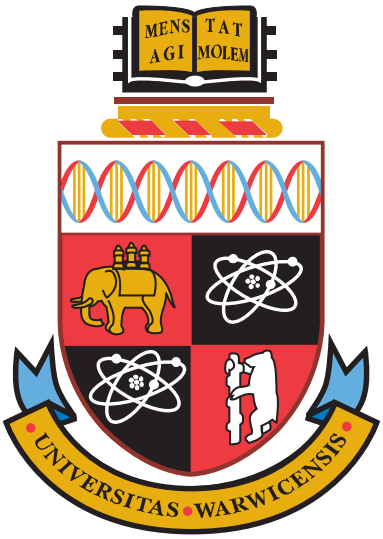
- If my thesis does include any substantial subsidiary material owned by third-party copyright holders, I have sought and obtained permission to include it in any version of my thesis available in digital format and that this permission encompasses the rights that I have granted to the University of Warwick and to the British Library.

5. LEGAL INFRINGEMENTS

I understand that neither the University of Warwick nor the British Library have any obligation to take legal action on behalf of myself, or other rights holders, in the event of infringement of intellectual property rights, breach of contract or of any other right, in the thesis.

Please sign this agreement and return it to the Graduate School Office when you submit your thesis.

Student's signature: .....  ..... Date: 05/08/2013



Multivariate Global Testing and Adaptive Designs

by

Giorgos Minas

Thesis

Submitted to the University of Warwick

in partial fulfilment of the requirements

for admission to the degree of

Doctor of Philosophy

Department of Statistics

May 2013

THE UNIVERSITY OF WARWICK

Contents

List of Tables

List of Figures

Acknowledgments

This thesis would not have come true without all those who have supported me in many different ways over the last three and a half years.

First, I would like to thank my supervisors Professor John A.D. Aston and Professor Nigel Stallard for the time and energy that they have spent with me these years to provide guidance and support in so many different ways. I could not imagine a better supervision. I would also like to thank Dr Fabio Rigat for the supervision during the first period of my PhD, and Tom E. Nichols for providing the fMRI data and his advice at various instances regarding fMRI applications.

Throughout these years I have had the opportunity to discuss my research with various members of the department. I would like to thank the people of the statistical inference and neuroimaging reading groups for many fruitful discussions. Special thanks go to the members of my review panels, Dr Elke Thonnes, Dr Sach Mukherjee and Professor Jane L. Hutton for providing useful feedback at various instances during my PhD.

I also feel grateful to the Department of Statistics at the University of Warwick for providing the facilities, the funding for attending various conferences, but most importantly the stimulating environment for research. I would also like to thank the Engineering and Physical Sciences Research Council that has funded my studies through the Warwick Centre for Analytical Science (EP/F034210/1).

During my PhD I have had the chance to make some new friends that I would like to thank for making my time more pleasant during these years. To name very few older ones, Chris Nam, Javier Rubio, Duy Pham, Hasinur Rahaman Khan, and

younger ones, Helen Ogden, Pantelis Samartsidis, Panayiota Touloupou, thank you. I would also like to thank Chris for proofreading this thesis and Pantelis, Habib Ganjgahi for helping me to produce an fMRI figure.

Lastly, I could not forget my friends, all my family members - my grandparents, my brothers-, sisters- and parents-in-law, my siblings, Giannis, Eleni and her family, and my parents, Kostas and Loukia - and especially my wife Kyriaki for their love, support and patience to see me finishing this thesis.

Declarations

I hereby declare that this thesis is based on my own research, except when stated otherwise, in accordance with the regulations of the University of Warwick, and has not been submitted elsewhere.

The methodology in chapter 4 and part of the results in chapter 8 are incorporated in a paper [Minas et al., 2012] published in *Statistics in Medicine* (2012, 31) with the title: "*A hybrid procedure for detecting Global Treatment Effects in Multivariate Clinical Trials: Theory and Applications to fMRI Studies*". This is joint work with F. Rigat, T.E. Nichols, J.A.D. Aston and N. Stallard.

The methods and results contained in chapters 6, 7 and 8 are joint work with J.A.D. Aston and N. Stallard. The methodological part of this work is submitted for publication in a paper entitled: "*Adaptive multivariate global testing*". A draft of this paper [Minas et al., 2013] is published in *CRISM research papers*. The application of this methodology to fMRI studies is incorporated in the conference paper entitled: "*ROI analysis of pharmafMRI data: an adaptive approach for global testing*" published in the "*Proceedings of 46th Scientific meeting of the Italian Statistical Society, 2012*". T.E. Nichols also participated in this work. A paper that considers wider application of this methodology to neuroimaging studies is to be submitted for publication.

In all of these publications I have taken the leading role both in terms of preparation of the material and conceptual work. In addition, all results presented in chapter 8 have been produced by me, except when otherwise indicated by references.

In addition to the work presented in this thesis, a contribution to the discussion of the paper: "*Group sequential tests for delayed responses*", by L.V. Hampson and C. Jennison, is published in the *Journal of the Royal Statistical Society Series B-Statistical Methodology*, (2013, 75). This is joint work with J.A.D. Aston.

Abstract

Global tests are a key research endpoint in multivariate studies. They provide an omnibus assessment of the overall effects across the multivariate outcomes. This global evaluation is clearly of high practical value in the field of neuroimaging, which has become increasingly important in recent years. Existing global testing methodologies, however, fail to accommodate the demands of neuroimaging studies that have typically small sample sizes and highly correlated local outcomes.

In this thesis a novel class of multivariate global tests is developed. The proposed tests are based on a formal framework for using prior information and accumulated data to learn the effect direction. This framework is used to construct test statistics that target the estimated effect direction, rather than the whole multivariate space, for detecting global effects. Adaptive designs are employed to allow for sequential modifications of the test statistics, based on accumulated data, without inflating the type I error.

A major focus in our methodology is power performance. The proposed tests are shown to be optimal in terms of predictive power. Furthermore, a power characterisation allowing us to explain the behaviour of our tests and perform simple power analysis is derived. An extensive power analysis, including comparisons to alternative global tests, is performed. Applications to neuroimaging studies are illustrated through two real examples. Our results show that the developed methodology can be particularly useful in cases where the sample sizes are small and prior information about the effect direction is available.

Abbreviations

MRI	Magnetic Resonance Imaging
fMRI	functional MRI
EEG	Electroencephalography
MEG	Magnetoencephalography
PET	Positron Emission Tomography
BOLD	Blood Oxygenation Level Dependent
CNS	Central Nervous System
GLM	General Linear Model
SPM	Statistical Parametric Mapping
ROI	Region(s) Of Interest
DFT	Discrete Fourier Transformation
OLS	Ordinary Least Squares
GLS	Generalised Least Squares
SS	Standardised Sum
PC	Principal Component
SSD	Single-Stage Design
GS	Group Sequential (design or test)
GSD	Group Sequential Design
AD	Adaptive design (or test)
CIP	Conditional Invariance Principle
CEF	Conditional Error Function
RSSR	Rate of Sample Size Reduction

Notation

The following notation is used throughout this thesis, unless otherwise stated. In addition to their statement here, they are usually described at their first occurrence. Unless otherwise stated, we use: (i) lower case, normal font type to represent scalars, (ii) lower case, **bold** font type to represent vectors and (iii) upper case, **bold** font type for matrices. No distinction is made between random variables and observed values in terms of notation, but, if not obvious by the context, this is made clear by appropriate descriptions.

$\mathbf{R}$	the set of real numbers
$\mathbf{R}^K$	the set of $K-$dimensional vectors with real entries
$\boldsymbol{z}^T$, $\boldsymbol{A}^T$	transpose of a vector $\boldsymbol{z}$, matrix $\boldsymbol{A}$
$\boldsymbol{c}_K$	$K-$dimensional vector with all entries equal to the scalar c
$ang(\boldsymbol{a}, \boldsymbol{b})$	angle, in measured degrees at the origin, between $\boldsymbol{a}$ and $\boldsymbol{b}$
$Diag(a_1, a_2, \ldots, a_K)$	$K \times K$ diagonal matrix with diagonal entries $a_1, a_2, \ldots, a_K$
$Diag(\boldsymbol{a})$	diagonal matrix with diagonal the vector $\boldsymbol{a}$
$\boldsymbol{I}_K$	$K \times K$ identity matrix
$E(\cdot)$	expectation function
$Pr(A)$	probability of event A
$Pr(A\|B)$	conditional probability of event A given event B

$N(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ
$N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$	$K-$dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$
t_ν	Student's t distribution with ν degrees of freedom
$t_\nu(\theta)$	Non-central t distribution with ν degrees of freedom and location/non-centrality parameter θ
$t_K(\nu, \boldsymbol{\theta}, \boldsymbol{S})$	$K-$dimensional non-central t distribution with ν degrees of freedom, location/non-centrality parameter $\boldsymbol{\theta}$ and scale matrix $\boldsymbol{S}$
$IW_{K\times K}\left(\nu, \boldsymbol{S}^{-1}\right)$	Inverse Wishart distribution with ν degrees of freedom and scale matrix $\boldsymbol{S}$
χ^2_K	χ^2 distribution with K degrees of freedom
$\chi^2_K(D)$	Non-central χ^2 distribution with K degrees of freedom and non-centrality parameter D
$T^2_{K,n}$	Hotelling's T^2 distribution with (K, n) degrees of freedom
$T^2_{K,n}(D)$	Non-central Hotelling's T^2 distribution with (K, n) degrees of freedom and non-centrality parameter D
$\phi(\cdot)$	The standard normal probability density function
$\Phi(\cdot)$	The standard normal cumulative density function
$\Psi_{\theta,\nu}$	The cumulative distribution function of $t_\nu(\theta)$
$\psi_{\theta,\nu}$	The probability density function of $t_\nu(\theta)$
Δ	Mahalanobis distance
k	dimension index
K	dimension of observations
i	subject index
n	sample size
j	stage index
J	maximum number of stages

H_0	null hypothesis
H_1	alternative hypothesis
α	nominal type I error
β	power
$1-\beta$	type II error
χ^2 test	single stage multivariate χ^2 test
T^2 test	single stage multivariate Hotelling's T^2 test
χ^2_{GS} test	group sequential test with test statistic the multivariate χ^2
T^2_{GS} test	group sequential test with test statistic the multivariate Hotelling's T^2 test
z test	single stage test with test statistic the linear combination $z-$ statistics with fixed weighting vector
t test	single stage test with test statistic the linear combination $t-$statistics with fixed weighting vector
z_{GS} test	group sequential test with test statistics the linear combination $z-$statistics with fixed weighting vectors
t_{GS} test	group sequential test with test statistics linear combination $z-$statistics with fixed weighting vectors
z^+ test	single stage linear combination $z-$test with weighting vector selected using prior information and/or pilot data
t^+ test	single stage linear combination $t-$test with weighting vector selected using prior information and/or pilot data
z^+_{AD} test	adaptive linear combination $z-$test with weighting vector initially selected using prior information and sequentially adapted to accumulating data at interim analyses
t^+_{AD} test	adaptive linear combination $t-$test with weighting vector initially selected using prior information and sequentially adapted to accumulating data at interim analyses

Chapter 1

Introduction

Multiple outcomes emerge in almost every area of scientific investigation. Naturally, when performing a scientific study, researchers wish to monitor a number of measures for each experimental unit. With more measures recorded and explored, one can develop greater knowledge about the problem. This multiplicity, however, does not solely arise from scientific curiosity and luxury to collect more information, but often because of the nature of the problem which implies that a single isolated measure is not sufficient to answer key experimental questions. Particularly, in medical studies, to evaluate the effects of a treatment on subjects, multiple symptoms and bodily functions need to be monitored. This situation occurs in many other fields of statistical application including industry, economics, ecology, biology and psychology, in cases where multiple events or phenomena are essential to evaluate an effect of interest.

A fundamental question arising in experimental studies, regardless of whether single or multiple outcomes are evaluated, is whether the results provide significant evidence for the existence of the effect of interest or not. This is most often a key question. In the presence of multiple outcomes, it is translated as whether the results provide significant evidence for an *overall* effect. An omnibus assessment throughout the multiple outcomes is then required. The global null hypothesis of

no treatment effect in any of the multiple outcomes is formed and suitable testing procedures are constructed. These *global tests*, as opposed to being interested in detecting "local" effects, as in effects on specific outcomes, simultaneously evaluate multiple outcomes to provide a global statement for the presence of the treatment effect.

Global testing is a classical field of statistical inference with several different approaches being available. Multiple testing methods, such as Bonferroni correction, can be used to evaluate global hypotheses. These are typically simple procedures, usually constructed under weak modeling assumptions, but they become conservative if the multiple outcomes are highly correlated. In the latter case, it is often useful to consider the multiple outcomes as multivariate observations and to incorporate correlations into the modeling assumptions. This multivariate approach is especially appropriate when the multiple outcomes are biologically (or in some other context) related. The classical multivariate global test, Hotelling's T^2, can efficiently detect effects in every direction of the multivariate space, when the sample size of the study, n, is sufficiently large. However, in settings where n approaches or becomes smaller than the observation dimension K, the T^2 test becomes respectively inefficient and inapplicable. This cost in efficiency, paid because of searching in every direction of the alternative space, seems particularly wasteful if prior knowledge about the direction of the effect is available.

The present thesis investigates global testing procedures in the presence of multivariate observations. This work is motivated by an area which introduces new challenges in global testing, highlighting some deficiencies of existing methodology and necessitating the demand for novel methodology. This is the very exciting, and rich in statistical applications, field of neuroimaging.

Neuroimaging uses powerful techniques, such as Magnetic Resonance Imaging (MRI), to explore the anatomy, function and pharmacology of the normal and abnormal brain. In neuroimaging studies, the global question whether the measures

of neural activity at various brain locations suggest a significant overall treatment effect is fundamental. These local neural measures, even after substantial summarisation, can have relatively large dimension. Furthermore, these measures are typically highly correlated and the effects across them are often dispersed in the sense that they are locally small, but, if combined, globally large. Another property of neuroimaging studies is that the great amount of research in the area and the spatial characterization of neural measures typically provide the researchers prior information about various aspects of their investigation. Finally, the high cost of neuroimaging equipment and expertise typically restricts sample size of neuroimaging studies to small levels.

These properties are taken into consideration in the present thesis to develop novel methodology for global testing. Multivariate assumptions are imposed on the observation vectors enabling us to incorporate correlations and combine dispersed local effects for a single global evaluation. The proposed tests are based on linear combinations of the observation vectors. The crucial element in this approach is the weighting vector reducing the observation vectors to scalar linear combinations. This defines the direction in which we decide to search for effects, and it can substantially affect both type I and type II error rates of the tests. A formal framework for selecting the weighting vector using prior information and pilot data without inflating the type I error is developed. This enables the proposed tests to attain high power levels for large sample sizes, but can be efficient even in situations where the sample size is limited to relatively low values.

In a major development of our methodology, global testing procedures are implemented within an adaptive design framework. Adaptive designs allow for interim design modifications, based on the observed data, without inflating the type I error rate. The use of adaptive design methodology increases the possible actions of the proposed procedures and can potentially improve efficiency. The global test statistics, initially constructed based on prior information, are sequential updated

based on accumulated data at subsequent interim analyses. Early termination of the study, due to early acceptance or rejection of the null hypothesis at interim analyses, is also possible within the adaptive design framework.

While the developed tests are analytically proved to control type I error, a major focus of our methodology is power performance. The test statistics of the constructed procedures are derived to be optimal, within the class of linear combination tests, with respect to predictive power given the information available at interim analyses. Furthermore, a framework for performing power analysis of linear combination tests is derived. In this, we reduce the complexities in performing power analysis of linear combination tests, by re-expressing the possibly high dimensional design space as a lower dimensional easily interpretable space, that is, still sufficient to determine power. These results provide wide understanding of the behaviour of linear combination tests and allow us to perform relatively simple power analysis. The main results of an extensive power analysis, including comparisons to alternative tests and application to neuroimaging studies, is provided in this thesis.

Finally, it is useful to note here that the methodology developed in this thesis is motivated by neuroimaging studies, but our framework is rather more generic and can be applied for multivariate global testing in many other fields. Biomedical studies and particularly clinical trials would likely provide an area of application. Clinical trials are studies undertaken to examine the effects of different medical interventions on human subjects [Friedman et al., 2010]. The issues of multiple outcomes, global testing, error rates control are often crucial in clinical trials and this is addressed by the present thesis.

1.1 Outline of the thesis

The remainder of this thesis is structured as follows.

In chapter 2, the motivating application of neuroimaging studies is intro-

duced. I briefly summarize the current state of the field, focusing on the two neuroimaging modalities, functional MRI (fMRI) and Electroencephalography (EEG), which I target in this work. fMRI and EEG data analysis is then discussed with special attention given to the types of preparatory analysis that generated the two datasets used throughout this thesis to illustrate applications of various global tests.

In chapter 3, we discuss the problem of testing global hypotheses. Various global tests available in the literature are discussed, with special attention given to their strengths and weaknesses. We first briefly discuss $p-$value adjustment methods, such as the Bonferroni global test. We then proceed to multivariate global tests which are the broad focus of this thesis. We discuss the fully multivariate χ^2 and Hotelling's T^2 tests and briefly introduce one-sided multivariate tests. Finally, we proceed to the class of linear combination tests which is the specific area of focus in this thesis. Here, we describe the available approaches in the literature and address the weaknesses which we attempt to mitigate using the methodology developed in later chapters.

In chapter 4, we develop novel methodology for performing linear combination tests. The class of linear combination tests is first formally introduced. Then, power-optimal linear combination tests are derived. Using this result, links to O'Brien linear combination and Hotelling's T^2 test are derived. The power-optimal linear combination tests use weighting vectors which depend on unknown model parameters. The weighting vectors which maximise predictive power given prior information and preliminary data obtained from a pilot study are then derived. The proposed z^+ and t^+ tests are discussed, while a comparison to the alternative approach of a fully Bayesian test is also provided.

Adaptive designs provide the possibility of substantially improving the tests developed in chapter 4. In chapter 5, the framework of adaptive design methodology is introduced. Here sequential designs are also discussed as they are strongly related, both conceptually and methodologically, to adaptive designs and they pro-

vide the framework for various alternative global tests considered for comparison to the developed methodology. The concepts that give rise to sequential and adaptive designs are first discussed. Group sequential and adaptive testing is next developed with the main attention given to the combination tests used in our methodology. Finally, we briefly describe various types of applications of these designs available in the literature, while the chapter is closed with a discussion for the current state, challenges and potentials of the field.

In chapter 6, we develop a methodology for performing adaptive linear combination tests. First, we formulate the $J-$stage tests with stage-wise statistics obtained via linear combinations. The power-optimal $J-$stage linear combination tests are then derived. The latter tests use weighting vectors which depend on the unknown modelling parameters. For practical implementation, a framework for sequentially updating the weighting vector, initially constructed based on prior information, using the data observed at interim analyses is constructed. Adaptation rules maximising the predictive power given the interim results are derived. These tests are analytically proved to control type I error.

The problem of performing power analysis of multivariate global tests and particularly linear combination tests is discussed in chapter 7. We derive a power characterisation of linear combination tests in terms of low-dimensional easily interpretable parameter summaries. The implications of these results, with respect to our understanding for linear combination tests and for performing power analysis, are discussed.

In chapter 8, the main results of an extensive power analysis are presented. We start by describing the effect of various design and model parameters on power performance. Comparisons between various global tests, including the constructed procedures, are next provided. Finally, application of various global tests on our real examples of an fMRI and an EEG study are considered.

The thesis finishes with a discussion of the main results as well as possible

future developments and extensions of the proposed methodology.

Appendix A provides some proofs of the results in chapter 6.

Chapter 2

Neuroimaging studies

2.1 Introduction

In this chapter, we attempt to outline the enormously exciting field of neuroimaging which has motivated the methodology developed in this thesis. Our target is to establish our motivations and to provide the necessary background to understand the examples to which we apply our methods.

We first briefly provide some background about neuroimaging and its current state. We then provide some more details on the neuroimaging modalities, fMRI and EEG, in which we are most interested. Here, we discuss fMRI and EEG data analysis and particularly Regions of Interest (ROI) analysis of fMRI data and frequency analysis of EEG data from which our real datasets are derived. We also introduce our real examples arising from a fMRI drug development study and an EEG depression study.

2.2 Neuroimaging

The study of the human brain has a long history tracing back at least to the father of western medicine Hippocrates. Since then, scientists from many disciplines including mathematics, physics, chemistry, biology, medicine and psychology studied

different aspects of the nervous system. In the late 1960s, scientists of all these disciplines decided to merge their knowledge under an interdisciplinary field coined *neuroscience*. The Society of Neuroscience, formed in 1971, set a common principal target for neuroscientists: to understand the structure and function of the normal and abnormal brain. This initiated an era of revolutionary achievements and great public interest with neuroscience being one of the leading areas of science today and the Society of Neuroscience being, according to Bear et al. [2007], the "largest and fastest-growing association of professional scientists in all experimental biology" [Bear et al., 2007; Squire et al., 2008].

Neuroimaging is the branch of neuroscience that uses various techniques to create images of the structure, function and/or pharmacology of the brain. Structural neuroimaging targets the description of brain anatomy, while functional neuroimaging attempts to describe the functional organization of the brain [Squire et al., 2008]. Technological development enabled the invention of various neuroimaging techniques over the 20th century [Raichle, 2000]. The consecutive discoveries of Positron Emission Tomography (PET) in 1980s and especially fMRI [Ogawa et al., 1992] in early 1990s led to an explosion of interest in functional neuroimaging over the last decades (see for example Friston [2009]).

Neuroimaging is now the predominant technique in behavioral and cognitive neuroscience [Cabeza et al., 2001; Friston, 2009] and it has a fast-growing role in psychiatry [Phillips, 2012], evidence-based neurology [Burneo et al., 2011] and image-guided neurosurgery [Aquilina et al., 2005]. An emerging application of neuroimaging is in the discovery and development of drugs for the treatment of disorders of the Central Nervous System (CNS). This is regarded by many authors as a great opportunity to respond to the increasing burden of CNS disorders by improving the efficiency of current practice in CNS drug development [Matthews et al., 2011; Wong et al., 2008].

Neuroimaging provides non-invasive and safe methods for rapid data acqui-

sition. Using the various modalities of neuroimaging, the living human brain can be studied with great spatio-temporal resolution while activated and possibly engaged to symptom-related tasks. Many authors have argued that neuroimaging can play an important role in early stages of drug development by providing objective markers of brain activity. Imaging biomarkers can replace subjective behavioral measures especially to support proof-of-concept studies and go/no-go decision making. At the current time, neuroimaging techniques are (strictly) not validated for drug development, but most major pharmaceutical companies are embracing this technology by establishing it in-house or via academic collaborations [Borsook et al., 2011; Wise and Tracey, 2006].

Nevertheless, there are several challenges to be overcome before these techniques become further established. The signal of various important neuroimaging modalities needs to be better understood and neuroimaging traits of CNS diseases need to be established. The high cost of the equipment and the need for trained individuals to run the experiments most often limits the sample size of the studies, while the pressure for efficiency remains high. Hence, there is a need for standardized statistical methodologies specialized to neuroimaging data analysis [Borsook et al., 2011; Whitcher and Matthews, 2006]. We respond to the latter problems, as we explain later, by providing methodology which uses the special properties of neuroimaging to achieve high efficiency even in settings with small sample sizes.

The modalities of functional neuroimaging can be broadly separated into two categories. The first, which mainly consists of Electroencephalography (EEG) and Magnetoencephalography (MEG), directly measures brain activity by capturing the electrical or magnetic signals produced by neurons during activation. The second, which is currently dominated by functional Magnetic Resonance Imaging (fMRI) and Positron Emission Tomography (PET), indirectly measures brain activity by capturing changes in local blood flow linked with neuronal activation. The nature of the various neuroimaging methods affect their spatial and temporal resolution

which along with their invasiveness and their cost are important properties in terms of application (see table 2.1).

Table 2.1: Properties of various neuroimaging modalities. Source: Lystad and Pollard [2009]

	PET	fMRI	EEG	MEG
Measure	indirect	indirect	direct	direct
Response	haemodynamic	haemodynamic	neuroelectrical	neuromagnetic
Invasive	yes	no	no	no
- confined	yes	yes	no	yes
- radiation	yes	none	none	none
Device cost	$8,000,000	$2,000,000	$100,000	$2,000,000
Operating cost	$1,500	$800	$150	$600
Temporal res	1-2 min	4-5 s	< 1 ms	< 1 ms
Spatial res	4 mm	2 mm	10 mm	5 mm

In the following, we discuss further two important neuroimaging techniques, fMRI and EEG, which motivate the methodology developed in this thesis.

2.3 fMRI

The most prominent form of fMRI is based on the so-called Blood Oxygenation Level Dependent (BOLD) contrast. BOLD fMRI (often simply called fMRI) uses the strong magnetic fields generated by MRI scanners to capture the local changes in blood oxygenation level that accompany neural activation. These local haemodynamic responses of the brain are recorded in 3−dimensional images with great spatial resolution (around $1-4$ millimeters). The temporal resolution (around $1-4$

seconds) of the scanner is restricted by various issues including the delay of the haemodynamic response to neural activation [Huettel et al., 2008].

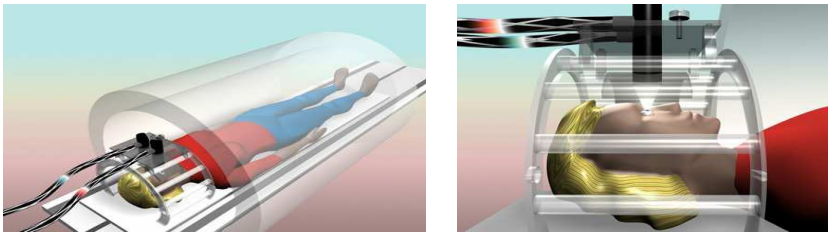


Figure 2.1: Schematic representation of a subject in a fMRI scan. Image by Duff Hendrickson, U.W., Copyright Hunter Hoffman, U.W.

The key properties of fMRI are the non-invasive, unharmful nature of the technique and the very high spatial resolution allowing for a study of the brain at the systems level. fMRI is currently one of the main drivers in understanding brain function. The potential of further application to drug development studies is discussed in a number of review publications (see Iannetti and Wise [2007]; Matthews et al. [2011]; Whitcher and Matthews [2006]; Wise and Tracey [2006] for overviews). fMRI is currently used to study the function and pharmacology of the brain under several pathologies, such as Alzheimer's disease [Pihlajamäki and Sperling, 2008] and drug addiction [Smith et al., 2010]). Honey and Bullmore [2004] report more than 50 published articles on psychopharmacological studies using fMRI while Schwarz et al. [2011a,b] provide guidelines for good imaging practice in pharmacological fMRI studies.

2.3.1 FMRI data analysis

The typical fMRI dataset produced by a single scanning session consists of BOLD recordings acquired from a small number of subjects (often around 15) during a relatively short period of time (few hundreds time points) at around $10^4 - 10^5$ voxels[1] throughout the brain.

[1]voxel: a three-dimensional volume element, with rectangular cuboid shape, analogous to the two-dimensional pixel.

The raw fMRI data is preprocessed using several techniques of signal processing, image processing and statistics. This typically involves realignment and spatial normalization using suitable transformations to register the raw data to a common reference image. Spatial smoothing and temporal filtering are also employed to eliminate the most common experimental artifacts (such as motion and scanner artifacts) and increase the signal-to-noise ratio [Poldrack et al., 2011]. The typical approach to model the preprocessed fMRI data is to apply mass-univariate General Linear Models (GLMs) to the time-series of each voxel separately. Normality is a common and generally acceptable assumption for the preprocessed fMRI data [Friston et al., 2007; Lindquist, 2008; Poldrack et al., 2011]. The default approach for statistical inference is based on maps of the brain, called Statistical Parametric Maps (SPMs), depicting the value of the t-statistics for each voxel. Parametric (for example random field theory) or non-parametric (for example permutation tests) approaches are then widely used to handle the huge multiple-testing problem of detecting activated voxels throughout the brain while controlling for false positives [Friston et al., 2007; Poldrack et al., 2011].

The burden of this multiple-testing problem can be alleviated by restricting the search for activated voxels only to specific brain areas. In the following, we briefly describe one method of data reduction. This is the ROI analysis applied to the fMRI study of the example in section 2.3.2.

ROI analysis

In fMRI data analysis, investigators often primarily target selected brain locations called regions of interest (ROI). ROI analysis is also very often reported as supplementary to the standard mass-univariate voxel-by-voxel analysis in fMRI studies. Compared with the voxel-by-voxel approach, ROI analysis provides a number of key benefits. First, in data exploration, it enables the investigator to take advantage of and find links with important findings that are currently available in the literature

regarding the anatomy and function of a great number of brain ROI. For similar reasons, ROI analysis is more suitable than voxel-by-voxel analysis for studying specific regional hypotheses about the drug action [Wise and Tracey, 2006]. Such hypotheses are more strict and therefore potentially more conformable to the regulations of drug authorities. Finally, ROI analysis results in a drastic reduction of data dimension and this is expected to substantially increase statistical power [Mitsis et al., 2007; Poldrack et al., 2011].

The first step in ROI analysis is to define the exact location of the ROI to be analyzed. This can be performed based on structural or functional features. Structural or anatomical ROI can be defined based on anatomical landmarks of the brain. These are described by standard brain atlases, such as the Talairach atlas [Talairach and Tournoux, 1988], which are widely available. In some cases, brain atlases can even be specific to the area of investigation. For example, Mitsis et al. [2007] use a "pain-atlas" for a pain treatment study. Based on such atlases, investigators define the detailed location of ROI in each subject either in terms of the subject's brain anatomy, derived using high-resolution structural MRI, or using "probabilistic atlases" reflecting the anatomical variability between subjects. On the other hand, functional ROI are often derived using an independent "localizer" scan to identify voxels in particular brain areas that show a characteristic response of interest. Finally, ROI can be defined based on previous studies, preferably from a meta-analysis on the domain of interest. It is important to note that to control the type I error, the ROI locations need to be defined prior to analyzing the data of the main study [Poldrack et al., 2011].

The next step in ROI analysis is to quantify the measure to be extracted from each ROI. One method is to count the activated voxels within the defined ROI. However, this approach, used mainly in early fMRI studies, is very sensitive to the specified activation threshold. More commonly, voxel-wise estimates, $\hat{\beta}$, of a chosen parameter, β, expressing an effect of interest, are extracted from the

mass-univariate GLM. These $\hat{\beta}$-values are often standardized to the design of the experiment by appropriate scaling. Summaries of the $\hat{\beta}$-values across the voxels of each ROI of each subject are then derived using either weighted (for example first Principal Component [Friston et al., 2007]) or un-weighted averages. These averages of the estimated effects constitute the ROI data which are used to derive statistical inference for the brain action at, between or across the selected ROI [Poldrack et al., 2011]. The main steps for deriving ROI data are also described in figure 2.2.

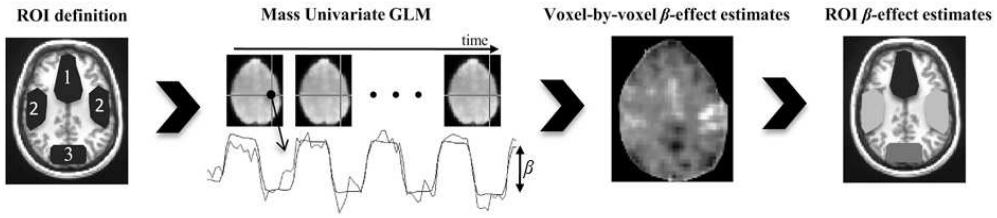


Figure 2.2: Typical steps of ROI analysis producing the multivariate outcome used in our real example in section 2.3.2. The preprocessed series of fMRI images are modelled at voxel-by-voxel resolution using mass univariate General Linear Models (GLMs). Suitable estimates of parameter values (β) expressing the treatment effect in each voxel are first extracted from the GLM and then averaged across the pre-defined ROI to produce a multivariate outcome where each component corresponds to a measure of the treatment effect in a specific ROI.

2.3.2 Example: fMRI drug development study

For the purposes of drug development, a fMRI study was conducted by Glaxo-SmithKline plc. A total of 13 subjects participated in the study. At the planning stage, the following anatomical ROI (see figure 2.3) were defined: 1. Anterior Cingulate (AC), 2. Atlas Amygdala (A), 3. Caudate (C), 4. Dorsolateral Prefrontal Cortex (DL), 5. Globus Pallidus (GP), 6. Insula (I), 7. Orbitofrontal cortex (OF), 8. Putamen (P), 9. Substantia Nigra (SN), 10. Thalamus (T), 11. Ventral Striatum (VS).

ROI summary data was extracted from the mass-univariate GLM applied to preprocessed fMRI data. For this, the voxel-wise $\hat{\beta}$ parameter estimates were aver-

aged across the voxels of each ROI for each subject. The available data represents the difference between paired treatment-placebo observations across ROI for each subject.

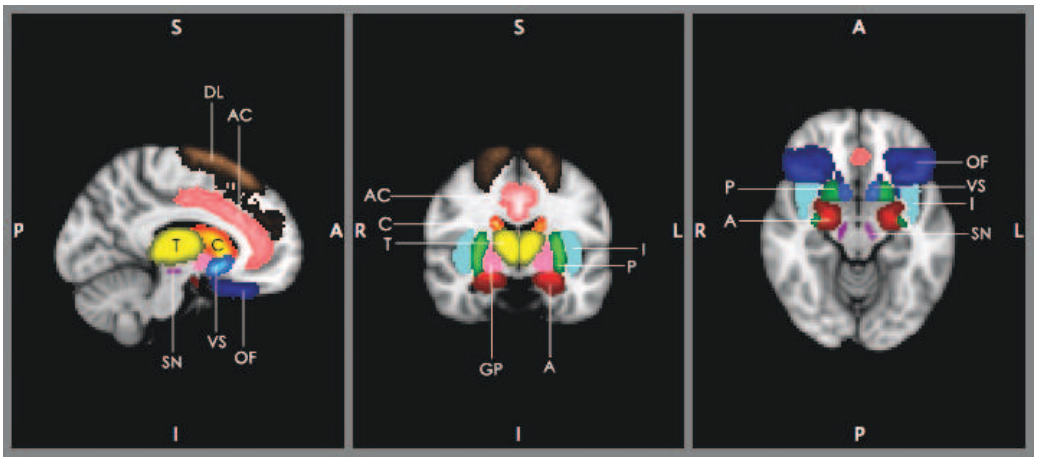


Figure 2.3: Approximate locations of ROI used in the fMRI study example. Each of the following ROI are indicated by a different colour and abbreviation: Anterior Cingulate (AC-pink), Atlas Amygdala (A-red), Caudate (C-red/yellow), Dorsolateral Prefrontal Cortex (DL-copper), Globus Pallidus (GP-magenta), Insula (I-purple), Orbitofrontal cortex (OF-blue), Putamen (P-green), Substantia Nigra (SN-cyan), Thalamus (T-yellow), Ventral Striatum (VS-blue/lightblue)

The means, standard deviations and correlations matrix of this ROI dataset are presented in table 2.2. As we can see, effect sizes differ across ROI, the standard deviations are relatively large and generally high correlations are observed.

Using this data, the investigators wished to address a fundamental question arising in neuroimaging studies. This concerns the existence of global treatment effects across ROI. The term *global* is used here to stress the difference to being interested in local effects, as in effects at specific ROI. The null hypothesis to be evaluated is then whether the treatment-placebo differences provide no statistically significant evidence for treatment effects across the selected ROI. To answer this global question, within this setting but also more generally, we employ various testing procedures introduced in the next chapters. We now turn our interest to another neuroimaging modality: electroencephalography (EEG).

Table 2.2: Means, standard deviations and correlations of ROI data ($n_T = 13$) of the fMRI study.

ROI (k)	AC	A	C	DL	GP	I	OF	P	SA	T	VS
$\bar{x}_k$	0.01	-0.06	0.08	0.08	0.14	0.02	0.08	0.06	0.10	0.10	0.13
s_k	0.33	0.33	0.17	0.22	0.33	0.28	0.36	0.39	0.32	0.33	0.32
$r_{AC,k}$	1	0.70	0.87	0.88	0.73	0.89	0.66	0.81	0.26	0.95	0.70
$r_{A,k}$		1	0.55	0.61	0.72	0.77	0.65	0.68	0.59	0.68	0.66
$r_{C,k}$			1	0.89	0.72	0.87	0.47	0.80	0.27	0.90	0.74
$r_{DL,k}$				1	0.71	0.76	0.73	0.77	0.27	0.87	0.62
$r_{GP,k}$					1	0.86	0.51	0.90	0.54	0.70	0.90
$r_{I,k}$						1	0.45	0.85	0.46	0.86	0.84
$r_{OF,k}$							1	0.44	0.09	0.65	0.39
$r_{P,k}$								1	0.49	0.82	0.89
$r_{SA,k}$									1	0.30	0.55
$r_{T,k}$										1	0.74
$r_{VS,k}$											1

2.4 EEG

EEG records the electrical activity of the brain. Specifically, EEG captures the summed electrical potential generated by synchronously activated neurons (tens of thousands). This activity is captured by electrodes typically placed at various locations of the scalp. EEG has unrivalled temporal resolution (around 1-10 milliseconds) with recordings regarded as "real-time" measures. On the other hand, spatial resolution has fundamental physical limits with up to a number of sites, often called channels, sampled in routine studies. For clinical studies, typically 19 channels are used, but if necessary 32, 64 or even more than 100 channels can be

used [Gevins, 2002].

The high temporal resolution, the relatively low cost compared to other neuroimaging techniques, as well as the non-invasive, safe, portable equipment are the main advantages of EEG. During the last decade, neuroscientists have shown great interest in combining EEG with fMRI to achieve high spatial and temporal resolution [Mulert and Lemieux, 2010]. There is also interest in the so-called event-related potentials (ERPs) which employ EEG to record brain response to specific stimuli [Handy, 2005]. EEG has made great contribution in studying the effects and pharmacology of various CNS-related disorders (see [Bauer and Bauer, 2005]), particularly epilepsy [Thompson and Ebersole, 1999] and depression [Steiger and Kimura, 2010]. An example of the latter application is described in section 2.4.2.

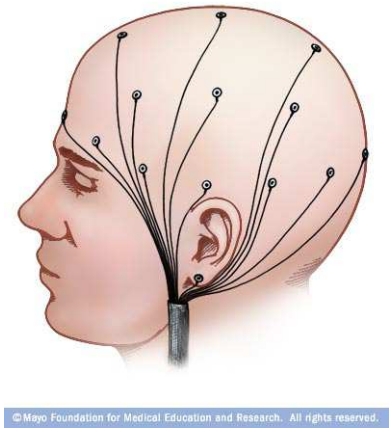


Figure 2.4: Schematic representation of the EEG electrodes placed on the scalp. Image source: www.riversideonline.com

2.4.1 EEG data analysis

EEG raw data of a single subject consists of multiple time-series recorded at a number of different location or channels. They are typically preprocessed by applying high- and low-pass filters to eliminate experimental artifacts and possibly signals of no interest for the study. Baseline correction is also applied by subtracting a "mean signal" from all time points to eliminate slow level shifts produced by experimental

artifacts such as sweating and muscle tension [Hauk, 2013; Sanei and Chambers, 2007].

EEG data can be analyzed as observed in the time-domain, but they are often transformed to the frequency-domain to perform data analysis. Frequency data analysis is particularly important in EEG. The rhythms that characterize normal and abnormal brain activity in EEG are categorized in four major frequency ranges. These are (from low to high frequency): delta ($0.5 - 4$ Hertz), theta ($4 - 8$ Hertz), alpha ($8 - 13$ Hertz) and beta ($13 - 30$ Hertz). In terms of cognition states, delta is associated with deep sleep and theta with creative inspiration and deep meditation. Alpha frequency is associated with relaxed awareness and beta is observed during active thinking, focus and creation. Abnormal EEG signal in alpha and beta frequency ranges is associated with various CNS related disorders (for example epilepsy [Sanei and Chambers, 2007]).

The Discrete Fourier Transformation (DFT) is typically used to transform the time-series of each channel to the frequency domain. Formally, the DFT is defined as

$$y(f) = \sum_{t=1}^{N} x(t) \exp^{-i2\pi ft},$$

where $x(t)$ the response at time t and $y(f)$ the amplitude of the spectrum at frequency f. The square of the amplitude, $(y(f))^2$, named the power spectrum, or more commonly its logarithm, $\log\left(y(f)^2\right)$, is then used in data analysis [Rao, 2010]. In practice, EEG data is typically segmented into short time intervals before applying the DFT. These epochs are defined by the clinicians, by taking into account the experimental design, to ensure that they have similar characteristics. To avoid discontinuities at the epoch edges, overlapping between them and suitable window functions (such as the Hanning or Gaussian window) are often used. DFT is then applied to the time-series of each epoch (after windowing and overlapping) to derive their power spectrum function. To increase the low signal-to-noise ratio of EEG

data, the power spectra are typically averaged across these epochs to derive a summary aggregate power spectrum (stationarity assumptions are made here) for each subject. This average power spectrum function, or the average power spectrum at specific frequencies, are then often used for data analysis [Kropotov, 2009; Sanei and Chambers, 2007]. Gaussianity is the typical assumption for the logarithms and other transformations of the average power spectra [Collura et al., 2009; Lopes da Silva, 2005; Sanei and Chambers, 2007].

2.4.2 Example: EEG depression study

Läuter et al. [1996] published frequency data from an EEG study. As they describe, the data is collected from $n = 19$ depressive patients at the beginning and at the end of a six week therapy. The published data represent the changes on the absolute[2] theta power of channels $3, 4, 5, 6, 7, 8, 17, 19$ ($K = 9$ locations, see figure 2.5) during the therapy of each patient.

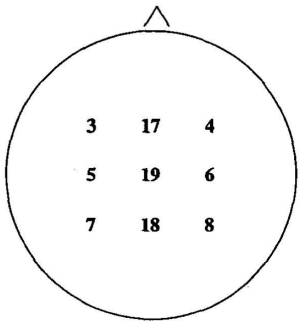


Figure 2.5: Schematic presentation of the position of the channels used in the EEG depression study

In table 2.3 we present the means, standard deviations and correlation matrix of the data. Note that the correlations especially of nearby locations are high. Furthermore, an increase in theta power is indicated in all channels, but the standard deviations are generally high.

[2]The term absolute is used in EEG data analysis to distinguish from the relative power where, in the latter, the absolute power is divided by the total power in all frequencies

Similarly to the fMRI study in section 2.3.2, the investigators were primarily interested in the global treatment effects. That is, they wished to investigate whether the differences in the absolute theta power suggest a statistically significant treatment effect across the selected channels. This question can be expressed in terms of a global hypothesis.

Table 2.3: Means, standard deviations and correlations for the EEG depression study presented in Läuter et al. [1996].

ch. (k)	3	4	5	6	7	8	17	18	19
$\bar{x}_k$	0.87	1.59	1.04	1.15	0.85	0.85	1.42	0.75	1.00
s_k	2.95	3.51	2.36	2.25	2.28	2.07	3.26	2.64	2.36
$r_{3,k}$	1	0.93	0.81	0.80	0.58	0.49	0.93	0.49	0.53
$r_{4,k}$		1	0.63	0.78	0.34	0.45	0.93	0.28	0.49
$r_{5,k}$			1	0.79	0.85	0.71	0.73	0.71	0.76
$r_{6,k}$				1	0.60	0.79	0.82	0.63	0.78
$r_{7,k}$					1	0.62	0.46	0.68	0.60
$r_{8,k}$						1	0.52	0.60	0.78
$r_{17,k}$							1	0.40	0.57
$r_{18,k}$								1	0.44
$r_{19,k}$									1

2.5 Conclusions

Neuroimaging is a very important and exciting field of neuroscience. Neuroimaging studies can provide great insight into the normal and abnormal brain anatomy and function. However, for these studies to become further established, particularly in a

clinical trials setting, a number of challenges remain outstanding and some of them regard statistical methodology.

Some of these challenges are illustrated by the real examples of the fMRI and EEG study. These studies share common properties which are typical in neuroimaging studies. Firstly, high correlations are observed especially but not exclusively between nearby locations. Secondly, the observed effects are dispersed across different locations, in the sense that they are locally small but, as we show in section 3.3.1, globally large. This suggests combining the local outcomes, rather than treating them separately, to detect global effects. In addition, due to the high cost of these studies, the sample size is small and even after the reduction in data dimensionality (for example using ROI or frequency analysis) remains close to the observations' dimension. Furthermore, for these studies but also more generally in neuroimaging, there are opportunities to elicit prior information from earlier results. In section 8.4 we provide some examples of such prior information which may arise from the spatial characterization of the signal and from the nature of the study. Lastly, as our examples illustrate, effects are often expressed in different directions between brain locations with hyperactivation simultaneously occurring with deactivation.

These properties of neuroimaging studies are taken into consideration in the methods introduced in later chapters. The deficiencies of existing global tests in application to neuroimaging are discussed and novel methodology is developed.

Chapter 3

Global testing

3.1 Introduction

As we discussed in the previous chapter, in neuroimaging studies the neural activity is simultaneously investigated at multiple brain locations. More generally, in biomedical studies but also in experiments performed in many other fields (for example industry, ecology and psychology), investigators are often interested in a number of outcomes and only rarely focus on a single measure. This situation often becomes necessary due to the nature of the research questions. For example, in order to evaluate the treatment effect on patients, multiple symptoms and relevant body functions need to be monitored and hence multiple outcomes need to be evaluated [D'Agostino and Russell, 2005; Dmitrienko et al., 2010].

In this chapter, we consider methods that can be used to evaluate treatment effects in settings where multiple outcomes are studied. Specifically, we are interested in testing procedures which can be used to evaluate global effects observed across these outcomes. That is, as opposed to focusing on local effects on specific outcomes, we consider methodology which evaluates multiple outcomes simultaneously to provide an omnibus assessment of the effects over all the outcomes. These methods address the fundamental question of whether or not, overall, the

observed outcomes suggest a significant treatment effect. [D'Agostino and Russell, 2005; O'Brien, 1984; Pocock, 1997; Sankoh et al., 1997]. This methodology can be used to evaluate the question posed by the examples of neuroimaging studies, as seen in sections 2.3.2 and 2.4.2, regarding global treatment effects across multiple brain locations.

In the following, we consider various testing procedures, focusing mainly on the essence of their methodology and discussing when they are appropriate. We discuss $p-$value adjustment methods and multivariate tests, including "fully" multivariate tests and the class of linear combination tests which is the main methodological focus in this thesis. We begin with a general formulation of the global testing problem.

Formulation

The $K-$dimensional observation vectors, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iK})^T$, of subjects $i = 1, 2, \ldots, n$, are assumed to be independent with common mean vector $E(\boldsymbol{x}) = \boldsymbol{\mu} = (\mu_1, \ldots, \mu_K)^T$ and covariance matrix (the symmetric positive definite) $\boldsymbol{\Sigma} = (\sigma_{kk'})_{k,k'=1}^{K}$. The mean vector $\boldsymbol{\mu}$ is often interpreted as the effect of interest or the treatment effect. We wish to test the global null hypothesis of no treatment effect against the two-sided alternative. That is,

$$H_0 : \boldsymbol{\mu} = \mathbf{0}_K = (0, 0, \ldots, 0)^T \qquad \text{versus} \qquad H_1 : \boldsymbol{\mu} \neq \mathbf{0}_K. \tag{3.1}$$

In global testing, and more generally in statistical hypothesis testing, we are interested in controlling or minimising the potential incorrect decisions, called errors, of the test. Specifically, we focus on controlling the type I error rate as in ensuring that

$$Pr(\text{ reject } H_0 \mid \boldsymbol{\mu} = \mathbf{0}_K) \leq \alpha, \tag{3.2}$$

and minimising the type II error rate

$$1 - \beta = Pr(\text{ not reject } H_0 \mid \boldsymbol{\mu} = \boldsymbol{\mu}_1), \qquad (\boldsymbol{\mu}_1 \neq \mathbf{0}_K) \tag{3.3}$$

or, equivalently, maximising the power of the test

$$\beta = Pr(\text{ reject } H_0 \mid \boldsymbol{\mu} = \boldsymbol{\mu}_1). \tag{3.4}$$

If the type I error is equal to the significance level α, that is, the equality in (3.2) is satisfied, we say that the test controls/maintains the type I error exactly or, simply, the test is exact. The target of maximising power is sometimes replaced by minimising the sample size of the test (or other design parameters) while controlling power at a fixed level. We stress that in this thesis power is denoted by the letter β rather than the (more common) expression $1 - \beta$, to simplify notation.

Note that the testing procedures which follow equally apply to the two-sample setting with common covariance matrix. This can be formulated in terms of two independent samples, $\boldsymbol{x}_{Ai}$, $\boldsymbol{x}_{Bi}$, $i = 1, 2, \ldots, n$ from groups A and B, with $E(\boldsymbol{x}_{Ai}) = \boldsymbol{\mu}_A$, $E(\boldsymbol{x}_{Bi}) = \boldsymbol{\mu}_B$, $var(\boldsymbol{x}_{Ai}) = var(\boldsymbol{x}_{Bi}) = \boldsymbol{\Sigma}$ which are used to test the null hypothesis $H_0 : \boldsymbol{\mu}_A - \boldsymbol{\mu}_B = \mathbf{0}_K$ against the two-sided alternative $H_0 : \boldsymbol{\mu}_A - \boldsymbol{\mu}_B \neq \mathbf{0}_K$. Furthermore, the setting of paired multivariate observations (as in observations before and after a treatment), where the multivariate outcomes are set equal to $\boldsymbol{d}_i = \boldsymbol{x}_{Ai} - \boldsymbol{x}_{Bi}$, $i = 1, 2, \ldots, n$, can also be accommodated by the methods which follow. Finally, with a few trivial changes in the procedures to follow, the situation where the hypotheses of interest are $H_0' : \boldsymbol{\mu} = \boldsymbol{\mu}_0$ and $H_1' : \boldsymbol{\mu} \neq \boldsymbol{\mu}_0$, with $\boldsymbol{\mu}_0 \neq \mathbf{0}_K$ can also be accommodated. However, to simplify notation we continue with the above one-sample presentation.

3.2 $P-$value adjustment methods

The above global hypotheses can be evaluated using $p-$value adjustment methods, such as the Bonferroni test. In the well-known Bonferroni correction test, the local outcomes, $x_1, x_2, \ldots, x_K$, are used to construct the statistics $t_1, t_2, \ldots, t_K$, with corresponding $p-$values $p_1, p_2, \ldots, p_K$, to test the local null hypotheses

$$H_{0k} \ : \ \mu_k = 0, \quad k = 1, 2, \ldots, K.$$

The Bonferroni test rejects the $k-$th local null hypothesis H_{0k} if and only if $p_k \leq \alpha/K$. Note that the global null hypothesis H_0 in (3.1) can be written as the intersection of the local null hypotheses, formally

$$H_0 \ = \ \bigcap_{k=1}^{K} H_{0k}, \tag{3.5}$$

and thus rejection of a local null hypothesis implies rejection of the global null hypothesis.

Thus, by ordering the $p-$values as $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(K)}$, we can write the global Bonferroni test as

$$\text{reject } H_0 \quad \text{iff} \quad p_{(1)} \leq \alpha/K. \tag{3.6}$$

Due to the first-order Bonferroni inequality (Boole's inequality) this procedure controls the type I error at the nominal α level [D'Agostino and Russell, 2005].

The global Bonferroni test is easy and convenient to apply and it does not require any distributional assumption to control the type I error. It can even be used if the observations of each dimension are measured at a different scale. On the other hand, the Bonferroni method relies entirely on the smallest $p-$value and can often be conservative (that is, type I error rate substantially lower than the

nominal α level) and inefficient (that is, power at unexpectedly low levels). This is the case when, as in our motivating example, high correlations exist between the local outcomes. The problem becomes worse as the dimension, K, of the observation vectors increases. Pocock et al. [1987] and Dmitrienko et al. [2010] found, through simulation studies, that for large positive correlations, especially when these are higher than 0.5, the type I error rate of the Bonferroni method is substantially lower than the nominal level. The results in Dmitrienko et al. [2010] show that the conservatism is considerably larger if the observation dimension is increased even from $K = 2$ to $K = 5$. We next consider the application of the global Bonferroni test to the examples in sections 2.3.2 and 2.4.2.

Example: Global Bonferroni test for fMRI and EEG study data

We compute the values of the t test statistics and the corresponding two-sided $p-$values (12 degrees of freedom) at each of the 11 ROI.

Table 3.1: The local $t-$ and $p-$values for the observations collected at each ROI used in the fMRI study.

ROI	AC	A	C	DL	GP	I	OFC	P	SA	T	VS
t_k	0.03	-0.72	1.80	1.28	1.55	0.21	0.76	0.59	1.14	1.10	1.46
p_k	0.98	0.48	0.10	0.23	0.15	0.84	0.46	0.57	0.28	0.29	0.17

The smallest $p-$value $p_{(1)} = p_3 = 0.10$ is clearly larger than $\alpha/K \approx 0.0045$ for $\alpha = 0.05$, $K = 11$ and thus the Bonferroni method fails to reject H_0.

Similarly, in the next table, we present the $t-$ and $p-$values of the theta frequency observations recorded at each channel used in the EEG depression study.

Table 3.2: The local $t-$ and $p-$values for the observations collected at each channel used in the EEG depression study.

ch.	3	4	5	6	7	8	17	18	19
t_k	1.29	1.97	1.92	2.22	1.63	1.80	1.90	1.24	1.84
p_k	0.22	0.06	0.07	0.04	0.12	0.09	0.07	0.23	0.08

The smallest $p-$value is $p_{(1)} = p_4 = 0.04 > \alpha/K \cong 0.0055$ for $\alpha = 0.05$, $K = 9$ and thus the Bonferroni method fails to reject H_0. □

A number of modifications of the Bonferroni method exist in the literature. Simes [1986] global test rejects H_0 if and only if $p_{(k)} \leq k\alpha/K$, for at least one k, $k = 1, 2, \ldots, K$. This test does not rely heavily on the smallest $p-$value and it is less conservative and more efficient than the Bonferroni method. Further, despite the slight increase in computation, it is still very easy and convenient to apply. However, Simes' global test does not always control the type I error. Simes [1986] analytically proved that his test controls type I error for independent outcomes, while, through simulations, he showed that the type I error is also controlled for specific correlation structures under various distributions including the multivariate normal. Hommel [1988] also proposed two $p-$value adjustment methods, which control the type I error and are less conservative than Bonferroni method but more conservative than Simes global test [D'Agostino and Russell, 2005].

The above methods completely ignore correlations and thus they all become conservative when correlations are high. Some $p-$value adjustment methods accounting for correlations exist in the literature (for example James [1991], random field theory [Friston et al., 2007] and non-parametric [Westfall and Young, 1993] methods), but these tend to require complex calculations and they often rely on assumptions for specific observation structures in order to be efficient and/or to

maintain the type I error.

More generally, any multiple testing method can be re-written as a global test, because rejection of a single local null hypothesis implies rejection of the global null hypothesis. However, multiple testing methods treat local outcomes as independent entities rather than as components of a multivariate observation and they focus on detecting one or some few distinct local effects. Hence, they are more appropriate when we are interested in assessing local rather than global effects [D'Agostino and Russell, 2005; Dmitrienko et al., 2010]. Since these characteristics do not fit with our motivating application, we are driven to multivariate tests considered next.

3.3 Multivariate tests

There are various reasons why a multivariate test might be more appropriate than $p-$value adjustment methods for evaluating global hypotheses. First, multivariate global tests control type I error without the burden of multiplicity adjustments. Secondly, multivariate tests make direct use of, rather than ignore, correlations by incorporating them into the multivariate assumptions used to construct global test statistics. This is particularly important in studies, such as those arising in neuroimaging, where high correlations between local responses are typically observed.

Further, rather than evaluating local effects separately, multivariate tests combine the effects from the correlated outcomes to obtain omnibus assessment of treatment effects. This is particularly meaningful in cases where the multiple outcomes are biologically related and especially if they are replications of the same measure at different locations, as in neuroimaging. In terms of efficiency, multivariate methods can be advantageous compared to $p-$value adjustment methods, if the effects are locally small but globally large if combined. The latter situation is illustrated by our neuroimaging examples.

Finally, these multivariate global tests are often constructed based on dis-

tances between multivariate observations and linear combinations of the multiple outcomes are fundamental [D'Agostino and Russell, 2005]. As Rencher [2002] nicely put it, the construction of these linear combinations "reveal more (compared to multiple univariate tests) about how the variables unite to reject the null hypothesis".

Multivariate normality assumption

The multivariate tests to follow are based on the assumption of multivariate normality. That is, the $K-$dimensional observation vectors $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{iK})^T$ of subjects $i = 1, 2, \ldots, n$, are assumed to be independent and identically distributed multivariate normal random variables,

$$\boldsymbol{x}_i \sim N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \quad i = 1, 2, \ldots, n, \quad \text{independent} \tag{3.7}$$

with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ (as above).

Multivariate normality is the central assumption for many multivariate statistical methods and there are various reasons for this [Anderson, 2003; Mardia et al., 1979; Rencher, 2002]. A practical advantage of this assumption is the mathematical tractability of the multivariate normal distribution. The probability density function of a random variable $\boldsymbol{x} \sim N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

$$f(\boldsymbol{x}) = (2\pi)^{-k/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\right), \tag{3.8}$$

is a direct generalisation of its univariate counterpart and is completely determined by its first two moments, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Several useful results can be easily proved for the multivariate normal distribution, including that the linear combinations $y = \boldsymbol{w}^T\boldsymbol{x}$, with $\boldsymbol{w} \neq \boldsymbol{0}_K$, is also normally distributed with

$$y \sim N(\boldsymbol{w}^T\boldsymbol{\mu}, \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}) \tag{3.9}$$

and that uncorrelated normal random variables are independent. Another reason for the dominance of normality in multivariate statistics is that, in contrast to the univariate case, there are considerably fewer established alternatives with generalizations of non-normal univariate distributions, such as Gamma and Poisson, not being widely accepted. Last, but not least, the central limit theorem for multivariate random variables asymptotically links, under some restrictions, any other multivariate distribution to the multivariate normal. Therefore, multivariate normality can often be considered as an approximation of other multivariate distributions [Mardia et al., 1979].

These reasons contribute in establishing multivariate normality as the typical assumption for modelling continuous observations in various multivariate settings. These include our motivating neuroimaging applications of fMRI [Friston et al., 2007; Lindquist, 2008; Poldrack et al., 2011] and EEG [Collura et al., 2009; Lopes da Silva, 2005; Sanei and Chambers, 2007]. In the methods to follow, multivariate normality assumption is used almost universally.

3.3.1 Fully multivariate tests

We now consider the fully multivariate global tests which, as opposed to $p-$value adjustment methods, are constructed by considering the multiple outcomes as multivariate observations.

The likelihood-ratio test for multivariate normally distributed observations as in (3.7) and $\mathbf{\Sigma}$ known is based on the statistic,

$$\chi^2 = n\bar{\boldsymbol{x}}^T\mathbf{\Sigma}^{-1}\bar{\boldsymbol{x}}, \tag{3.10}$$

where

$$\bar{\boldsymbol{x}} = n^{-1}\sum_{i=1}^{n}\boldsymbol{x}_i \tag{3.11}$$

is the sample mean of the observation vector $\boldsymbol{x}$. The χ^2 statistic follows the non-

central $\chi^2_K(\Delta^2)$ distribution with K degrees of freedom, and non-centrality parameter

$$\Delta^2 = (\boldsymbol{\mu} - \boldsymbol{\mu}_0)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \boldsymbol{\mu}_0), \quad \text{here } \boldsymbol{\mu}_0 = \mathbf{0}_K, \tag{3.12}$$

where Δ is called the Mahalanobis distance of the vector $\boldsymbol{\mu}$ from $\boldsymbol{\mu}_0$ for covariance matrix $\boldsymbol{\Sigma}$ [Mahalanobis, 1930]. This measures expresses the standardized, with respect to $\boldsymbol{\Sigma}$, distance between a value, $\boldsymbol{\mu}$, of the mean vector and the hypothetical null mean vector, $\boldsymbol{\mu}_0$. Mahalanobis is the central distance in multivariate statistics. It generalizes the univariate signal to noise ratio μ/σ and can be interpreted as a measure of deviation from the null hypothesis. In the medical setting, it is a well known global measure of the strength of the treatment effect.

Under H_0, $\Delta = 0$ and χ^2 is (centrally) χ^2_K distributed with K degrees of freedom. The χ^2 test rejects H_0 if and only if

$$\chi^2 > \chi^2_{K,\alpha}, \tag{3.13}$$

where $\chi^2_{K,\alpha}$ is the $100(1-\alpha)$ percentile of the null χ^2_K distribution. The χ^2 test

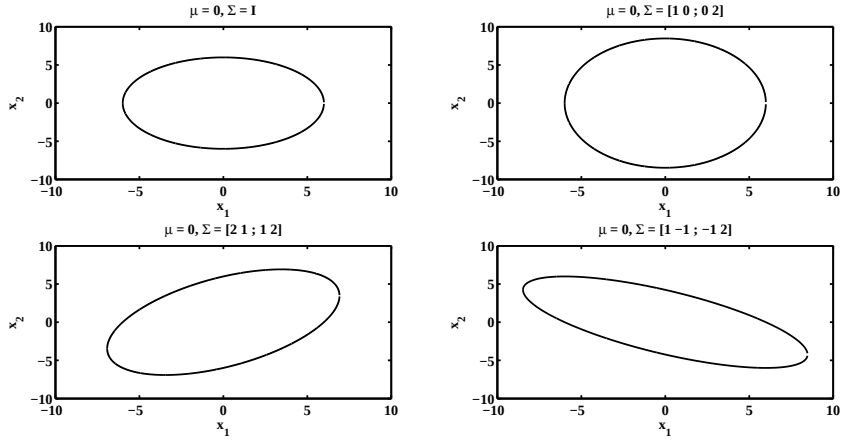


Figure 3.1: The ellipsoids $\chi^2 = \chi^2_{K,\alpha}$ of the two-dimensional ($K = 2$) χ^2 test for various $\boldsymbol{\Sigma}$. The outside area of the ellipsoids represents the rejection region of the χ^2 test. The direction in which the null hypothesis H_0 is rejected more easily, i.e. for smaller $\bar{\boldsymbol{x}}$, changes for different $\boldsymbol{\Sigma}$.

controls the type I error exactly. As we show in figure 3.1, the rejection region of the χ^2 test is the region outside the ellipsoid $n\bar{\boldsymbol{x}}\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}} = \chi^2_{K,\alpha}$ with the zero vector as the center. The size and shape of the ellipsoid is sensitive to the covariance structure in $\boldsymbol{\Sigma}$. As we can easily understand from figure 3.1, not only the size, but also the direction of the effect determines the value of the test statistic and thus whether H_0 is rejected or accepted (not rejected).

The main issue with the χ^2 test is that it requires the $\boldsymbol{\Sigma}$ matrix to be known. As the latter is most often not true, Hotelling's T^2 test, which takes $\boldsymbol{\Sigma}$ as unknown, is of much greater practical use. This is based on the statistic

$$T^2 = n\bar{\boldsymbol{x}}^T \boldsymbol{S}_{\boldsymbol{x}}^{-1}\bar{\boldsymbol{x}}, \tag{3.14}$$

where $\boldsymbol{S}_{\boldsymbol{x}}$ is the sample covariance matrix of $\boldsymbol{x}$, that is,

$$\boldsymbol{S}_{\boldsymbol{x}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T. \tag{3.15}$$

The test statistic T^2 is the generalisation of the (squared) univariate t test. It is the product of the sample size and the squared sample Mahalanobis distance

$$\bar{\boldsymbol{x}}\boldsymbol{S}_{\boldsymbol{x}}^{-1}\bar{\boldsymbol{x}},$$

which estimates the size of the global treatment effect.

Hotelling [1931] computed the distribution of T^2, the non-central $T^2_{K,n-1}(\Delta^2)$ distribution with non-centrality parameter Δ^2. The distribution is indexed by the dimension K and the degrees of freedom $n-1$. Under H_0, the non-centrality parameter is equal to 0 and T^2 follows the $T^2_{K,n}$ (central) distribution. The T^2 test rejects H_0 if and only if

$$T^2 > T^2_{K,n-1,\alpha} \tag{3.16}$$

where $T^2_{K,n-1,\alpha}$ is the $100(1-\alpha)$ percentile of the null $T^2_{K,n-1}$ distribution.

Hotelling's T^2 is the likelihood-ratio test for multivariate normal samples. It is invariant to affine transformations, $\tilde{\boldsymbol{X}} = \boldsymbol{C}\boldsymbol{X} + \boldsymbol{b}$, of the observation matrix $\boldsymbol{X}$, where $\boldsymbol{C}$ is a non-singular constant $K \times K$ matrix and $\boldsymbol{b}$ a constant $K \times 1$ vector. It is also admissible for testing H_0 under the multivariate normality assumption [Anderson, 2003]. It provides exact control over type I error and it is found to be robust against non-normality and heterogeneity between covariance matrices in two-sample tests [Rencher and Christensen, 2012].

A few more important properties of T^2 test arise from the critical values of T^2. First, T^2 requires $n - 1 > K$ to be applicable (otherwise $\boldsymbol{S}_{\boldsymbol{x}}$ is singular). Further, the T^2 distribution relates to the F distribution by the following equation

$$T^2_{K,n-1}(\Delta^2) = \frac{(n-1)K}{n-K} F_{K,n-K}(\Delta^2). \tag{3.17}$$

In the univariate case ($K = 1$), $t^2_{n-1} = F_{1,n-1} = T^2_{1,n-1}$, while asymptotically, as $n \to \infty$, $T^2_{K,\infty} = \chi^2_K$. As K increases, larger sample sizes are required for T^2 to approach χ^2. As shown in figure 3.2, for fixed sample sizes n, the critical value

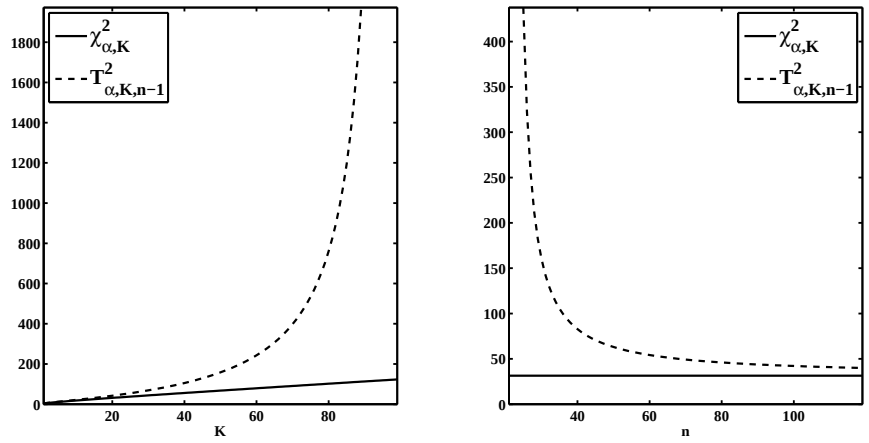


Figure 3.2: The critical values of the χ^2 and T^2 tests versus the dimension K (left panel) and sample size n (right). The critical values of T^2 can be much larger than χ^2 test for K approaching n. Here, $\alpha = 0.05$, $n = 100$ (left), $K = 20$ (right).

$T^2_{K,n-1,\alpha}$ increase exponentially with K. In particular, the critical values are extremely high for n close to K, a situation which is common in our motivating field of neuroimaging.

Example: Hotelling's T^2 test for the fMRI and EEG study data

For both the ROI data of the fMRI study and the theta frequency data of the EEG study, the sample Mahalanobis distance

$$D_{\boldsymbol{x}} = \left(\bar{\boldsymbol{x}} \boldsymbol{S}_{\boldsymbol{x}}^{-1} \bar{\boldsymbol{x}}\right)^{1/2}$$

is relatively large. Specifically, the sample Mahalanobis distance for the ROI data is $D_{\boldsymbol{x}_{ROI}} = 2.7657$ and for the frequency data $D_{\boldsymbol{x}_{THETA}} = 1.1537$. This suggests a large global treatment effect, contrary to the small local treatment effects, in both studies, which resulted in a failure to reject H_0 using the global Bonferroni test.

The T^2 statistic in (3.14) takes the value $T^2_{ROI} = 99.4413$, for the ROI data, and $T^2_{THETA} = 25.2881$ for the theta frequency data. Despite the large Mahalanobis distance, the T^2 test fails to reject H_0 in both studies ($p_{T^2_{ROI}} = 0.4656$, $p_{T^2_{THETA}} = 0.2486$), since the critical values are very high $T^2_{11,12,0.05} = 1280.7$, $T^2_{9,18,0.05} = 48.9302$. □

The χ^2 and Hotelling's T^2 tests are designed to detect deviations from the null hypothesis in any direction. Before we proceed to our main focus which are the linear combination tests, to provide a more complete picture of global testing, we briefly discuss in the next section another class of global tests, which target the detection of effects restricted to be positive in every dimension. Note that, as Rencher and Christensen [2012] suggests, the methodology for turning univariate two-sided tests to one-sided tests does not readily generalize to the multivariate setting, mainly due to the complexities introduced by the dependencies between the

multiple outcomes.

3.3.2 One-sided tests

One-sided tests evaluate the global null hypothesis against one-side alternatives, typically expressed as

$$H_1^+ : \mu_k \geq 0,\ k = 1, 2, \ldots, K,\ \text{(strict inequality for at least one } k). \tag{3.18}$$

In clinical trials, the alternative H_1^+ is often found appropriate to describe improvements expressed by increased treatment effects [Tang et al., 1993]. For evaluating H_0 against H_1^+, Kudo [1963] and Perlman [1969] derive the likelihood ratio tests (under multivariate normality) for covariance matrix assumed known and unknown, respectively. However, both of these tests require extremely challenging computations with their critical values depending on the covariance structure. The latter feature makes them practically infeasible even for small K. Their computational complexity arises from their test statistics' construction. For $\boldsymbol{\Sigma}$ known, this requires deriving an optimal projection, $\bar{\boldsymbol{x}}_+^*$, of the sample mean vector, $\bar{\boldsymbol{x}}$, in the positive orthant $\mathbf{R}_+^K = \{\boldsymbol{v} = (v_1, v_2, \ldots, v_K)^T \in \mathbf{R}^K | v_k \geq 0,\ k = 1, 2, \ldots, K\}$ of real numbers. Here optimality is defined in terms of maximisation of the sample Mahalanobis distance, $\bar{\boldsymbol{x}}_+^T \boldsymbol{\Sigma}^{-1} \bar{\boldsymbol{x}}_+$, of the projection, $\bar{\boldsymbol{x}}_+$. This optimization problem is substantially complicated by the dependencies across the multivariate outcomes.

Under independence, the latter optimal projection is derived by simply setting the negative entries of $\bar{\boldsymbol{x}}$ to zero. Tang et al. [1989a] use the latter result, by applying the one-sided likelihood ratio tests to observation vectors which are previously transformed using a matrix $\boldsymbol{A}$, such that $\boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{\Sigma}^{-1}$, to correct for correlations. This Approximate Likelihood Ratio test makes computations less complex and allows for critical values tabulation. However, it relies heavily on the non-unique definition of the matrix $\boldsymbol{A}$ which can be challenging, especially for unknown $\boldsymbol{\Sigma}$. Tang

et al. [1993, 1989a] proposed methods for deriving appropriate choices of matrix $\boldsymbol{A}$, but this requires intensive computations, particularly when the number of outcomes is high [D'Agostino and Russell, 2005].

Follmann [1996] proposes a simple test which also targets the one-sided problem. Follmann's test, often called T_+^2, rejects H_0 in favor of H_1^+ if both the sum of the mean vector is positive, that is, $\sum_{k=1}^{K} \bar{x}_k > 0$, and Hotelling's T^2 test rejects H_0 at twice the value of the significance level, 2α. Follmann [1996] proves type I error control of this test and showed its power robustness under various positive directions of the mean vector. Although, the dependence of T_+^2 to Hotelling's T^2 test implies that the former preserves the restriction for $n \gg K$ and therefore T_+^2, as T^2, is also often inappropriate for neuroimaging applications.

3.3.3 Linear Combination Tests

The fully multivariate tests considered in the previous section "look" in every direction of the two-sided or one-sided alternative space. Effects in every direction of the alternative space can be efficiently detected in settings where the sample size is sufficiently large. However, in settings where the sample size is restricted to be close to K, these methods become inefficient, while if the sample size is smaller than K fully multivariate tests are inapplicable. This efficiency cost from looking in every direction, seems avoidable and wasteful if prior knowledge about the direction of the effect is available. In this case, it seems more appropriate to construct a suitable test which targets this specific direction.

This statistical motivation leads to the development of "directional" or "linear combination" tests. The latter characterization addresses the construction of these tests, which are based on the linear combination $y = \boldsymbol{w}^T\boldsymbol{x}$, combining the observation vector, $\boldsymbol{x}$, using the weighting vector $\boldsymbol{w} = (w_1, w_2, ..., w_K)^T \neq \boldsymbol{0}_K$. In other words, the multivariate observation, $\boldsymbol{x}$, is reduced to the scalar linear combination, y, using the weighting vector $\boldsymbol{w}$. The geometrical interpretation of y is

particularly useful as we emphasize in chapter 7. The linear combination y can be seen as the magnitude of the projection of the observation vector $\boldsymbol{x}$ into the direction $\boldsymbol{w}$.

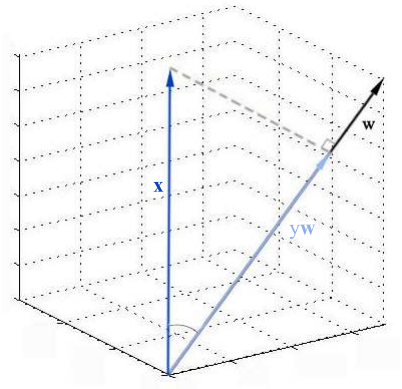


Figure 3.3: The observation vector $\boldsymbol{x}$, the weighting vector $\boldsymbol{w}$ and the projection vector $y\boldsymbol{w}$ with magnitude the linear combination y.

The linear combinations, $y_i = \boldsymbol{w}^T\boldsymbol{x}_i$, $i = 1, 2, \ldots, n$, of each subject's observation vector are treated as univariate outcomes to construct the standard z and t statistics

$$z = \frac{\bar{y}}{\sigma/\sqrt{n}}, \qquad t = \frac{\bar{y}}{s/\sqrt{n}},$$

for testing the null hypothesis H_0, when $\boldsymbol{\Sigma}$ is either known or unknown, respectively. Here, $\sigma^2 = \boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}$ is the variance and

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{3.19}$$

the sample mean and sample variance of the linear combination y, respectively.

The z and t statistics are scale invariant with respect to $\boldsymbol{w}$, that is, $z_{\boldsymbol{w}} = z_{c\boldsymbol{w}}$, $t_{\boldsymbol{w}} = t_{c\boldsymbol{w}}$ for any $c \neq 0$. Also, the distributions of both test statistics are invariant to scale transformations $\boldsymbol{x} \to c\boldsymbol{x}$ $(c \neq 0)$ of the observation vector $\boldsymbol{x}$. However, these tests are not generally invariant to affine transformation of the weighting vector or the responses. The latter properties emphasize the importance of the direction of

the weighting vector which is the crucial element in linear combination tests defining the region in which we search for treatment effects.

The selection of the weighting vector can substantially affect both type II and, if chosen based on observed data, type I error rate of the z and t test. This becomes clear if we consider the distribution of the z and t statistics. Under H_0, for fixed weighting vector $\boldsymbol{w}$, the z and t statistics are respectively standard normally $N(0,1)$ and Student's t distributed t_{n-1} with $n-1$ degrees of freedom. Thus the decision rules which reject H_0 if and only if

$$|z| > z_{\alpha/2}, \quad |t| > t_{n-1,\alpha/2},$$

where $z_{\alpha/2}$ and $t_{n-1,\alpha/2}$ are respectively the $100(1-\alpha/2)$ percentiles of the standard normal and of the Student's t_{n-1} distribution, specify hypothesis tests of size α. However, if the weighting vector depends on the observations (see examples below), the z and t statistics are not necessarily normally and t distributed and thus the type I error of the above tests can be inflated.

Furthermore, to see how crucial the weighting vector is for the power of the test, consider the distribution of the z and t statistics under H_1. Under H_1, the z and t statistics are normally and non-centrally t distributed with location parameter $\sqrt{n}\mu/\sigma$, where $\mu = \boldsymbol{w}^T\boldsymbol{\mu}$. The location parameter of the test statistics and thus power is largely affected by the direction of $\boldsymbol{w}$. For example, if $\boldsymbol{w}$ is fixed orthogonally to $\boldsymbol{\mu}$, the location parameter is zero and the power of the test is equal to α. It is thus important to ensure that $\boldsymbol{w}$ is not far from the effect direction and in particular not orthogonal to $\boldsymbol{\mu}$. These crucial issues are discussed next, but they are also thoroughly studied in chapters 4, 6, 7 and 8.

Here we attempt to outline the existing methodology in the class of linear combination tests. This can be partitioned into two main approaches. The first approach stems from the seminal work of O'Brien [1984] while the latter from Läuter

[1996]. We describe the methodology of these tests as well as some of their modifications and discuss their strengths and weaknesses. The latter discussion introduces the motivations behind the methodology described in later chapters.

O'Brien tests

O'Brien's [1984] seminal paper, as Dmitrienko et al. [2010] suggest, was a turning point, not only for linear combination tests, but for multiple testing procedures in general. In this paper, O'Brien derived several parametric and non-parametric tests for multiple endpoints settings in clinical trials. The parametric tests, developed under multivariate normality, are based on the assumption that the treatment effect is of equal strength across the multiple outcomes. That is, the mean vector $\boldsymbol{\mu}$ is assumed to be proportional to the uniform vector $\mathbf{1}_K = (1, 1, \ldots, 1)^T$. Based on this assumption, O'Brien derived global tests using Ordinary Least Squares (OLS) and Generalized Least Squares (GLS) methods. These can be expressed as linear combination tests with weighting vectors $\boldsymbol{w}_{OLS} = \mathbf{1}_K$, $\boldsymbol{w}_{GLS,1} = (\sigma_{11}^{-1}, \sigma_{22}^{-1}, \ldots, \sigma_{KK}^{-1})$ and $\boldsymbol{w}_{GLS,2} = \boldsymbol{\Sigma}^{-1}\mathbf{1}_K$, respectively.

O'Brien's tests are easily interpretable. The linear combination in OLS test is simply the sum of the multiple outcomes. In the first GLS test, the multiple outcomes are standardized to their standard deviation while in the second GLS test the weights are the column sums of the matrix $\boldsymbol{\Sigma}^{-1}$ and therefore outcomes that are less variant and less correlated (to other) outcomes receive greater weights.

The OLS z and t tests have fixed weighting vector and therefore, under multivariate normality, they exactly control the type I error. The GLS z test also control type I error, but for $\boldsymbol{\Sigma}$ unknown, O'Brien [1984] suggests replacing it with its sample estimate $\boldsymbol{S_x}$ and therefore the t statistic is no longer t distributed. Frick [1997] and Logan and Tamhane [2004] show that the type I error of the GLS tests can be inflated especially for small sample sizes and they obtain empirical critical values for type I error control.

O'Brien's tests, as we show in later chapters, are power optimal as in they attain maximum power, within the class of linear combination tests, if the assumption of uniform mean vector holds (additional assumptions are necessary for the OLS and the first GLS test). However, if the effect slips from the uniform direction, they can lose great amounts of power. D'Agostino and Russell [2005] suggest that O'Brien's tests are inefficient in situations where nonzero treatment effects are expected to occur in only a few endpoints. Furthermore, Tang et al. [1993] suggest that it might be difficult to know in advance whether the effect direction is close to the uniform direction.

O'Brien's methodology has been extended in many directions, including combination with the closure principle [Marcus et al., 1976] to provide multiple testing procedures [Lehmacher et al., 1991] and to develop consonant procedures [Bittman et al., 2009]. Practical application in various settings have also been considered [Dallow et al., 2008; Hemmelmann et al., 2004; Leroux et al., 2005]. Pocock et al. [1987] and Tang et al. [1993, 1989b] provide a useful discussion on O'Brien tests and extend their use respectively to asymptotically normal test statistics (for example survival and binary data) and group sequential designs. The wider class of group sequential linear combination tests with fixed weighting vectors is studied in later chapters.

Laüter exact tests

Läuter [1996] introduce a class of linear combination tests which targets high-dimensional settings being applicable even in situations where $n < K$. The proposed tests allow for determining the weighting vector based on the observed data without inflating the type I error. In particular, Läuter [1996] proves that, if the weighting vector is uniquely determined from the $K \times K$ sums of the products matrix $\boldsymbol{X}\boldsymbol{X}^T$, where $\boldsymbol{X} = [\boldsymbol{x}_1 \boldsymbol{x}_2 \dots \boldsymbol{x}_n]$, the type I error of the linear combination t test is exactly controlled. Note that, as Läuter [1996] confirms, the sample covariance matrix $\boldsymbol{S}_{\boldsymbol{x}}$

cannot be used instead of $\boldsymbol{X}\boldsymbol{X}^T = \boldsymbol{S}_{\boldsymbol{x}} + n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T$, since this may substantially inflate the type I error.

The type I error control under this weighting selection rule is proved using the theory of spherical matrix distributions [Fang and Zhang, 1990]. It is based on the invariance of the matrix $\boldsymbol{X}\boldsymbol{X}^T$ and the null distribution of $\boldsymbol{X}$ to orthogonal transformations. These invariance properties imply right-sphericity, under H_0, of both $\boldsymbol{X}$ and the linear combinations with weighting vector selected under the above rule which in turn implies that the linear combination $t-$statistic is exactly Student's $t-$distributed. Using these results and the theory of spherical matrix distributions, exact control of the type I error of the linear combination t test is deduced [Läuter et al., 1996].

Läuter et al. [1996] propose two methods for deriving the weighting vector from the matrix $\boldsymbol{X}\boldsymbol{X}^T$. The Standardized Sum (SS) test has weights equal to the inverse square root of the diagonal entries of $\boldsymbol{X}\boldsymbol{X}^T$, that is, $w_{SS,k} = \left(\sum_{i=1}^{n} x_{ik}^2\right)^{-1/2}$, $k = 1, ..., K$. Note that $\sum_{i=1}^{n} x_{ik}^2$ is an unbiased estimator of $n(\sigma_k^2 + \mu_k^2)$. Thus, correlations between the components of the response are not taken into account and components with larger means are down-weighted which contradicts the intuition for selecting weights. On the other hand, components with smaller variance receive larger weights which intuitively is a desirable property. As Läuter et al. [1996] suggest, the SS test is efficient "if the expected deviations from the hypothesis H_0 have the same direction for all K variables". In this case, $\boldsymbol{\mu} = \mathbf{1}_K$, and thus $\boldsymbol{w}_{SS}$ is expected to avoid down-weighting of outcomes with higher μ_k's. It might be useful to note the similarities with O'Brien's tests which are also powerful under the same restrictions.

The second test proposed by Läuter et al. [1996] is the Principal Component (PC) test which has weighting vector $\boldsymbol{w}_{PC}$ equal to the solution of the eigenvalue

problem

$$(\boldsymbol{X}\boldsymbol{X}^T)\boldsymbol{w} = \lambda Diag(\boldsymbol{X}\boldsymbol{X}^T)\boldsymbol{w}, \qquad \boldsymbol{w}^T Diag(\boldsymbol{X}\boldsymbol{X}^T)\boldsymbol{w} = 1,$$

corresponding to the largest eigenvalue λ. Here, $Diag(\boldsymbol{X}\boldsymbol{X}^T)$ is a diagonal matrix with main diagonal entries as in $\boldsymbol{X}\boldsymbol{X}^T$. The weighting vector $\boldsymbol{w}_{PC}$ can be alternatively written as $\boldsymbol{w}_{PC} = Diag(\boldsymbol{X}\boldsymbol{X}^T)^{-1/2}\tilde{\boldsymbol{w}}$, where $\tilde{\boldsymbol{w}} = (\tilde{w}_1, \tilde{w}_2, \ldots, \tilde{w}_K)^T$ is the solution of the eigenvalue problem

$$Diag(\boldsymbol{X}\boldsymbol{X}^T)^{-1/2}\boldsymbol{X}\boldsymbol{X}^T Diag(\boldsymbol{X}\boldsymbol{X}^T)^{-1/2}\tilde{\boldsymbol{w}} = \lambda\tilde{\boldsymbol{w}}.$$

The latter presentation shows that the weighting vector $\boldsymbol{w}_{PC}$ is the standardized, with respect to $Diag(\boldsymbol{X}\boldsymbol{X}^T)^{-1/2}$, first eigenvector of the matrix

$$Diag(\boldsymbol{X}\boldsymbol{X}^T)^{-1/2}\boldsymbol{X}\boldsymbol{X}^T Diag(\boldsymbol{X}\boldsymbol{X}^T)^{-1/2}.$$

This also implies the connection between the weights of the SS and PC tests, that is, $w_{PC,k} = \tilde{w}_k w_{SS,k}$, $k = 1, 2, ..., K$, which indicates that under certain conditions they have similar behaviour. The PC tests, similarly to the SS test, is appropriate if uniform effects are expected across the multiple outcomes. Läuter et al. [1996] argue that the PC test can be useful under the one factorial structure between the mean vector and the covariance matrix, as in $\boldsymbol{\Sigma} = \boldsymbol{\Psi} + c^2\boldsymbol{\mu}\boldsymbol{\mu}^T$, with $\boldsymbol{\Psi}$ being a diagonal matrix of nuisance variances and c is a constant. However, the task of deriving analytical or empirical results supporting these arguments remain outstanding and this is partly due to the complex construction and thus interpretation of the $\boldsymbol{w}_{PC}$ vector.

Despite these difficulties, Läuter's methodology is appealing mainly for two reasons. The first is that it remains applicable for sample sizes smaller than K. The second is that it provides a framework for deriving the weights using the observed

data without inflating the type I error. The proposed class of tests can be used for purposes beyond global testing (see for example Läuter et al. [2009]) and they have been applied in various high-dimensional settings [Adolf et al., 2011; Almeida and Ledberg, 2002].

Example: OLS, SS and PC t tests for fMRI and EEG study data

We apply O'Brien's OLS and Läuter's SS and PC t tests to the fMRI and EEG data. The GLS t tests are not considered since they do not control the type I error for small sample sizes. The OLS statistic is equal to $t_{ROI,OLS} = 0.9329$, for the ROI data, and $t_{THETA,OLS} = 2.1118$, for the theta frequency data. For the ROI data, $t_{ROI,OLS}$ is smaller to the critical value $t_{12,0.05/2} = 2.1788$ and thus OLS does not reject H_0 ($p_{ROI,OLS} = 0.3693$). In comparison, for the theta frequency data, $t_{THETA,OLS} = 2.1118$ which is larger than the critical value $t_{18,0.05/2} = 2.1009$ and thus OLS (narrowly) rejects H_0 ($p_{THETA,OLS} = 0.0489$).

For the SS and PC t tests, we obtain $t_{ROI,SS} = 0.9552$ and $t_{ROI,PC} = 0.9421$ which are both smaller than $t_{12,0.05/2} = 2.1788$ and thus the SS and PC do not reject H_0 ($p_{ROI,SS} = 0.3583$, $p_{ROI,PC} = 0.3647$). This is in contrast to the tests for the theta frequency data where, as with the OLS test, the $t-$values, $t_{THETA,SS} = 2.112$ and $t_{THETA,PC} = 2.114$, are both larger than the critical value $t_{18,0.05/2} = 2.1009$ and thus they also (narrowly) reject H_0 ($p_{SS} = 0.0489$, $p_{PC} = 0.0487$). Note that OLS, SS and PC tests have similar $t-$ and $p-$values. □

3.4 Conclusions

It is clear that no universally optimal solution can be found for global testing. Bonferroni-type methods can be more appropriate in the presence of one or few significant independent outcomes. Multivariate tests are more appropriate when the multiple outcomes are highly correlated and the effects are locally small but

globally large. Driven by our motivating application of neuroimaging studies, we mainly focus on multivariate tests.

If the sample size is sufficiently large, particularly substantially larger than the observations' dimension, Hotelling's T^2 test is very efficient. However, if the sample size is small and prior knowledge about the direction of the effect is available, linear combination tests provide an appealing alternative solution.

D'Agostino and Russell [2005] address three issues arising with O'Brien-type linear combination tests. The first considers the inflation of type I error in GLS tests. The second issue regards the inappropriateness of these tests in cases where the prior knowledge about the direction of the effect is either not reliable or it indicates a non-uniform effect direction. The third issue is concerned with situations where the second GLS test gives negative weights, but also more generally with the interpretability of results derived under O'Brien-type tests.

Except from the first issue regarding type I error control, which is addressed by Läuter's exact tests, the latter two issues remain outstanding. As we discuss later, the methodology developed in later chapters attempts to address these issues and provide satisfactory solutions.

Chapter 4

Optimal linear combination tests

4.1 Introduction

As we discussed in the previous chapter, D'Agostino and Russell [2005] raise three issues regarding O'Brien-type linear combination tests. These are in fact directly connected to the most important targets not only for O'Brien-type linear combination tests, but generally every testing procedure.

The first target is type I error control. This is the main target in hypothesis testing, especially within the context of clinical trials, in order to provide credibility to "positive" outcomes, that is, rejection of H_0. The second is type II error minimisation or equivalently power maximisation. This is required to ensure that "negative" outcomes, i.e. not rejection of H_0, are not derived simply because the study has insufficient power (for example insufficient sample size), but also to ensure that sufficient power is attained efficiently, i.e. without excessive use of the human or monetary resources. The third target is interpretability of the derived conclusions. This is often required because investigators are not satisfied by a single statement whether the null hypothesis is rejected or not, but they wish to be able to explain

the outcome. The linear combination tests developed in this chapter attempt to provide methodology satisfying these targets.

The proposed linear combination tests explicitly use prior knowledge to select the weighting vector $\boldsymbol{w}$. The use of prior knowledge for selecting the weighting vector has been suggested earlier by various authors. Specifically, Läuter et al. [1996]; Pocock et al. [1987] and Tang et al. [1993, 1989b] addressed that O'Brien's approach to select $\boldsymbol{w}$, under the assumption of uniform effect structure, reflects the investigators prior knowledge about the direction of the effect. They suggested that, if prior knowledge suggests a different structure due to different clinical or effect importance across outcomes, this should be expressed in the weighting vector. Furthermore, in our motivating application of neuroimaging studies, there is typically at least some vague prior knowledge about the effect structure. This can arise, for instance, from the spatial structure of the multivariate outcome. The methods developed next provide a formal framework for optimally incorporating such prior information into weighting vector selection. However, prior information is used only to select the weighting vectors and not for the final data evaluation.

For selecting the weighting vector, in addition to prior information, we also suggest using data from a preliminary sample obtained from a pilot study. Pilot studies are often performed prior to the main study to derive information, such as estimates of nuisance parameters, which is then used to decide various aspects of the design of the main study [Lancaster et al., 2004]. Birkett and Day [1994] and Shih et al. [2004] discuss the use of pilot studies for determining the sample size of the main study. Gupta and Perlman [1974] use a preliminary sample to decide whether to include additional, possibly costly or time-consuming, variables to perform a Hotelling's T^2 test. Westfall et al. [1998] use pilot data to select the component-wise significance levels in multiple testing. The use of pilot data, in our methods, allows the possibility of re-assessing prior information for the effect direction and potentially correcting the weighting vector.

The weighting vectors are optimally selected to maximise the predictive power of the test based on prior information and the pilot data. Predictive power is the chance to reject the null hypothesis H_0, at the final analysis, given the information available at a certain stage. Specifically, predictive power is defined as the probability to reject H_0 given an information set $\mathcal{I}$,

$$b = Pr(\text{ reject } H_0 \mid \mathcal{I}).$$

The information set $\mathcal{I}$ may include prior information and/or observed data. Note that predictive power is a probability defined in the Bayesian context. It can be derived by averaging the rejection probability over the distribution of the parameters given the information set $\mathcal{I}$ and for this reason it is often called expected or average power.

Predictive power has been used for sample size calculation [Huson, 2009; Lan and Wittes, 2012; O'Hagan and Stevens, 2001; Spiegelhalter and Freedman, 1986], monitoring clinical trials and stochastic curtailment [Spiegelhalter et al., 1986], treatment selection [Kimani et al., 2009], population selection [Brannath et al., 2009] and to select the component-wise significance levels in multiple testing [Westfall et al., 1998]. The use of predictive power for sample size calculation and stochastic curtailment has been criticised for incorporating the whole distribution of the parameters imposing probabilities for even extreme values. Thus, if investigators consider predictive power in the same scale with the power of the test, this may unnecessarily lead to large sample sizes and early stopping [Jennison and Turnbull, 1990]. However, as becomes clearer in later sections, these issues do not apply to our methods. In our methods, predictive power is used to allow for incorporating, in a natural and plausible way, prior information and pilot data in determining the weighting vector. By selecting the weighting vectors that maximise predictive power, our selection becomes power-optimal with respect to the collected information.

This chapter is organised as follows. We start by formulating linear combination tests and discussing some of their principal characteristics. We stress the importance of the weighting vector and proceed to derive the weighting vector maximising the power of the test. We then derive corollaries of the latter result which provide links to alternative global tests. The optimal weighting vector depends on the unknown model parameters and thus we then proceed to derive the predictive power-optimal weighting vectors used in the proposed tests. The chapter is closed with discussion of some of the properties of the latter tests.

4.2 Formulation

We now formally set-up the problem of evaluating the global hypotheses using linear combination tests of multivariate observations. Linear combination tests have already been briefly formulated and discussed in the previous section, however, to ensure clarity of our results, the problem is considered again from the beginning.

Let the $K-$dimensional response vectors $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{iK})^T$ for subjects $i = 1, 2, \ldots, n$ be independent Gaussian random variables

$$\boldsymbol{x}_i \sim N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{4.1}$$

with mean $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_K)'$ and covariance matrix $\boldsymbol{\Sigma} = (\sigma_{kk'})_{k,k'=1}^K$. Here, the diagonal entries of the positive definite matrix $\boldsymbol{\Sigma}$ are the variances, $\sigma_{kk} = \sigma_k^2$, of the local outcomes and the off-diagonal entries the covariances, $\sigma_{kk'} = \sigma_k \sigma_{k'} \rho_{kk'}$, where $\rho_{kk'}$ the correlations between the outcomes x_k and $x_{k'}$, $k, k' = 1, 2, \ldots, K$, $k \neq k'$.

We wish to test the global null hypothesis of no treatment effect against the two-sided alternative

$$H_0 : \boldsymbol{\mu} = \mathbf{0}_K = (0, 0, \ldots, 0)^T \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \mathbf{0}_K. \tag{4.2}$$

The multivariate observations are reduced to the scalar linear combinations

$$y_i = \boldsymbol{w}^T \boldsymbol{x}_i = \sum_{k=1}^{K} w_k x_{ik}, \quad i = 1, 2, \ldots, n, \tag{4.3}$$

with non-zero fixed weighting vector $\boldsymbol{w} = (w_1, w_2, ..., w_K)^T \neq \mathbf{0}_K$. The linear combinations y_i for subjects $i = 1, 2, \ldots, n$, are independent Gaussian random variables

$$y_i \sim N(\mu, \sigma^2), \tag{4.4}$$

where $\mu = \boldsymbol{w}^T \boldsymbol{\mu}$ and $\sigma^2 = \boldsymbol{w}^T \boldsymbol{\Sigma} \boldsymbol{w}$. Note that, under H_0, $y_i \sim N(0, \sigma^2)$, $i = 1, 2, \ldots, n$.

The likelihood-ratio z and t statistics for testing H_0 against H_1 on the basis of the random sample y_i, $i = 1, ..., n$ are

$$z = \frac{\bar{y}}{\sigma/\sqrt{n}}, \qquad t = \frac{\bar{y}}{s/\sqrt{n}}, \tag{4.5}$$

for σ known and unknown, respectively. Here,

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i, \qquad s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{4.6}$$

are the sample mean and sample variance of the linear combinations.

Under H_0, the z and t statistics are standard normal, $z \sim N(0, 1)$, and Student's t distributed with $n - 1$ degrees of freedom, $t \sim t_{n-1}$. Thus, the decision rule to reject the null hypothesis H_0 if and only if respectively

$$|z| > z_{\alpha/2}, \qquad |t| > t_{n-1,\alpha/2}, \tag{4.7}$$

specify hypothesis tests of size α. It might be useful to stress that the latter tests, henceforth called linear combination z and t tests or simply z and t tests, exactly control the type I error at the nominal α level.

For any arbitrary value of $\boldsymbol{\mu}$, the z and t statistics are respectively normally and non-centrally t distributed, $z \sim N(\bar{\theta}, 1)$ and $t \sim t_{n-1}(\bar{\theta})$ with location parameter

$$\bar{\theta} = \theta\sqrt{n}, \qquad \theta = \mu/\sigma = \frac{\boldsymbol{w}^T\boldsymbol{\mu}}{\left(\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}\right)^{1/2}}. \tag{4.8}$$

The latter parameter θ, often interpreted as the signal-to-noise ratio, is central in the methods to follow. The power of these tests, for given values $\boldsymbol{\mu} \neq \mathbf{0}_K$ and $\boldsymbol{\Sigma}$, are respectively

$$\beta_z = Pr\left(|z| > z_{\alpha/2}\right), \qquad \beta_t = Pr\left(|t| > t_{n-1,\alpha/2}\right). \tag{4.9}$$

Note that, unlike Hotelling's T^2 test, the linear combination z and t statistic do not depend on the dimension of the response K and hence they are applicable even for $n < K$. However, as we explained in the previous chapter, this is attained by restricting our search for treatment effects to a single direction defined by the weighting vector. In what follows, we focus on deriving an optimal weighting vector maximizing the power of the herein described linear combination z and t tests.

4.3 The power-optimal z^* and t^* tests

Power maximisation is one of the most important targets in hypothesis testing. It is an intuitively correct approach since it targets maximization of the probability to detect a treatment effect. We wish to derive the weighting vector maximizing the power of linear combination z and t tests. The next lemma is the first step for computing this power-optimal weighting vector.

Lemma 4.3.1. *Under* (4.1)*, the power,* β_z*, of the single-stage z test is non-decreasing in the absolute value of* θ *in* (4.8)*.*

Proof. The power of the single-stage z test can be written as

$$\beta_z = Pr\left(|z| > z_{\alpha/2}\right) = 1 - \Phi\left(z_{\alpha/2} - \bar{\theta}\right) + \Phi\left(-z_{\alpha/2} - \bar{\theta}\right),$$

where $\Phi(\cdot)$ the cumulative distribution functions of the standard normal distribution $N(0,1)$. Its derivative with respect to θ is

$$\frac{\mathrm{d}\beta_z}{\mathrm{d}\theta} = \sqrt{n}\left(\phi\left(z_{\alpha/2} - \bar{\theta}\right) - \phi\left(-z_{\alpha/2} - \bar{\theta}\right)\right),$$

where $\phi(\cdot)$ the density of the standard normal distribution. The latter derivative is non-negative for $\theta \geq 0$ and negative for $\theta < 0$. Therefore, β_z is non-decreasing in $|\theta|$. □

The above result can be generalized to the t test. For this we use the well-known representation of the cumulative distribution function, $\Psi_{\theta,\nu}(\cdot)$, of the non-central $t_\nu(\theta)$ distribution (see for example Johnson et al. [1995])

$$\Psi_{\theta,\nu}(t) = \int_0^\infty \Phi\left(\frac{xt}{\sqrt{\nu}} - \theta\right) h_\nu(x)\, \mathrm{d}x, \tag{4.10}$$

where $h_\nu(x)$ the density of the χ_ν distribution with ν degrees of freedom.

The following result is proved by showing that for every fixed value in the support of the χ_ν distribution ($\nu = n - 1$), similar results to the z test can be obtained.

Lemma 4.3.2. *Under* (4.1)*, the power,* β_t*, of the single-stage* t *test is non-decreasing in the absolute value of* θ *in* (4.8)*.*

Proof. The power, β_t, of the single-stage t test can be written as

$$\beta_t = Pr\left(|t| > t_{\nu,\alpha/2}\right) = 1 - \Psi_{\bar{\theta},\nu}\left(t_{\nu,\alpha/2}\right) + \Psi_{\bar{\theta},\nu}\left(-t_{\nu,\alpha/2}\right).$$

Using equation (4.10), we derive that

$$\frac{\mathrm{d}\beta_t}{\mathrm{d}\theta} = \int_0^\infty \sqrt{\nu}\left(\phi\left(\frac{xt_{\nu,\alpha/2}}{\sqrt{\nu}} - \bar{\theta}\right) - \phi\left(\frac{xt_{\nu,\alpha/2}}{\sqrt{\nu}} + \bar{\theta}\right)\right) h_\nu(x)\mathrm{d}x.$$

For any $x \geq 0$, the quantity $\phi\left(\frac{xt_{\nu,\alpha/2}}{\sqrt{\nu}} - \bar{\theta}\right) - \phi\left(\frac{xt_{\nu,\alpha/2}}{\sqrt{\nu}} + \bar{\theta}\right)$ is nonnegative for $\theta \geq 0$ and negative for $\theta < 0$. Thus, the derivative $\frac{\mathrm{d}\beta_t}{\mathrm{d}\theta}$ is nonnegative for $\theta \geq 0$ and negative for $\theta < 0$, which implies that β_t is non-decreasing in $|\theta|$. □

Lemma 4.3.1 and 4.3.2 imply that the power of the single-stage z and t tests is maximized if and only if the value of θ is maximized. These results, in addition to being crucial for deriving theorem 4.3.1 below, can also be useful for more general settings, as for instance in linear regression, to construct power-optimal procedures.

Considering the linear combination z and t tests, lemma 4.3.1 and 4.3.2 imply that in order to maximize the power of these tests with respect to the weighting vectors $\boldsymbol{w}$, it is sufficient to maximize the value of θ. Using this result, we next derive the power-optimal weighting vector.

Theorem 4.3.1. *Under* (4.1)*, the power of the single-stage z and t tests is maximized with respect to $\boldsymbol{w}$ if and only if $\boldsymbol{w}$ is proportional to*

$$\boldsymbol{\omega}^* = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}. \tag{4.11}$$

Proof. The Generalized Cauchy-Schwarz inequality (see page 178 [Anderson, 2003]) states that

$$\theta^2 = \frac{(\boldsymbol{w}'\boldsymbol{\mu})^2}{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}} \leq \boldsymbol{\mu}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu},$$

with equality obtained if $\boldsymbol{w} = \boldsymbol{\omega}^*$ in (4.11). The results then follow by lemmas 4.3.1 and 4.3.2. □

The optimal weighting vector, $\boldsymbol{\omega}^*$, in (4.11) is a standardization of the treatment effect $\boldsymbol{\mu}$, with respect to $\boldsymbol{\Sigma}$. The form of $\boldsymbol{\omega}^*$ implies that larger local effects μ_k,

smaller local variances σ_k^2 and lower correlation values $\rho_{kk'}$, $k' \neq k$, all contribute to larger ω_k^*. We write z^* and t^* the optimal z and t tests with weighting vector $\boldsymbol{w} = \boldsymbol{\omega}^*$ and their power functions β_z^* and β_t^*.

Using theorem 4.3.1, we can derive simple results describing mean and covariance structures for which O'Brien's OLS and GLS tests [O'Brien, 1984] are power-optimal.

Corollary 4.3.1. *The power-optimal* $\boldsymbol{\omega}^*$ *in* (4.11) *is equal to*

1. *the weighting vector* $\boldsymbol{w}_{GLS,2} = \boldsymbol{\Sigma}^{-1}\mathbf{1}_K$, *if* $\boldsymbol{\mu} \propto \mathbf{1}_K$,
2. *the weighting vector* $\boldsymbol{w}_{GLS,1} = (\sigma_1^{-1}, \sigma_2^{-1}, \ldots, \sigma_K^{-1})^T$, *if* $\boldsymbol{\mu} \propto \mathbf{1}_K$ *and* $\boldsymbol{\Sigma}$ *is a diagonal matrix, as in the correlations* $\rho_{kk'} = 0$, $\forall\ k, k' = 1, 2, \ldots, K$, $k \neq k'$,
3. *the weighting vector* $\boldsymbol{w}_{OLS} = \mathbf{1}_K$, *if* $\boldsymbol{\mu} \propto \mathbf{1}_K$ *and* $\boldsymbol{\Sigma}^{-1}\mathbf{1}_K = c\mathbf{1}_K$, c *constant, that is,* $\mathbf{1}_K$ *is an eigenvector of* $\boldsymbol{\Sigma}^{-1}$ *with* c *the corresponding eigenvalue.*

Proof. The result is proved by applying the appropriate values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ at $\boldsymbol{\omega}^*$. □

From theorem 4.3.1, we can also derive the following result allowing us to compare the optimal power, β_t^*, of the t test to the power of Hotelling's T^2 test β_{T^2}.

Corollary 4.3.2. *For fixed values of the parameters* $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ *and sample size* $n \in (K, \infty)$, *where* $K > 1$,

$$\beta_t^* > \beta_{T^2}.$$

Proof. Under (4.1), the t^* statistic follows the non-central t distribution with $n-1$ degrees of freedom and non-centrality parameter $\theta^*\sqrt{n}$. Here,

$$\theta^* = \frac{\boldsymbol{\omega}^{*T}\boldsymbol{\mu}}{\left(\boldsymbol{\omega}^{*T}\boldsymbol{\Sigma}\boldsymbol{\omega}^*\right)^{1/2}} = \frac{\boldsymbol{\mu}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}}{(\boldsymbol{\mu}^T\Sigma^{-1}\Sigma\Sigma^{-1}\boldsymbol{\mu})} = \left(\boldsymbol{\mu}^T\Sigma^{-1}\boldsymbol{\mu}\right)^{1/2} = \Delta,$$

where Δ is the Mahalanobis distance. Hence, the square of the t^* statistic, t^{*2}, follows the non-central $F-$distribution with non-centrality parameter Δ^2 and $(1, n-$

1) degrees of freedom, that is, $t^{*2} \sim F_{1,n-1}(\Delta^2)$. Thus, the power of the t^* test can be written as

$$\beta_t^* = Pr\left(t^{*2} > t_{n-1,\alpha/2}^2\right) = Pr\left(F_{1,n-1}(\Delta^2) > F_{1,n-1,\alpha}\right),$$

where $F_{1,n-1,\alpha}$ is the $100(1-\alpha)$ percentile of the $F_{1,n-1}$ distribution. Further, the power of Hotelling's T^2, using the relation between the T^2 and F distributions in (3.17), can be written as

$$\beta_{T^2} = Pr\left(F_{K,n-K}(\Delta^2) > F_{K,n-K,\alpha}\right).$$

The result then follows by corollary 2.1 in Gupta and Perlman [1974] which implies the inequality

$$\beta_t^* = Pr\left(F_{1,n-1}(\Delta^2) > F_{1,n-1,\alpha}\right) > Pr\left(F_{K,n-K}(\Delta^2) > F_{K,n-K,\alpha}\right) = \beta_{T^2}.$$

□

Note that the sample size restriction, in corollary 4.3.2, is required only to ensure that Hotelling's T^2 test is applicable. For smaller sample sizes, the T^2 test is not applicable. The above result stress the scope for improvement to the power of Hotelling's T^2 test by using linear combination tests. It shows that the power of the t^* test is strictly larger than the power of T^2, while, as we can see in Figure 4.1, the difference between the two power functions can be considerably high, especially for small sample sizes.

Although these results are promising, it still remains that the optimal weighting vector $\boldsymbol{\omega}^*$ depends on the unknown modelling parameters. Therefore, the optimal weighting vector is unknown and cannot be used in practice. We next propose a method to optimally select the weighting vector in practice.

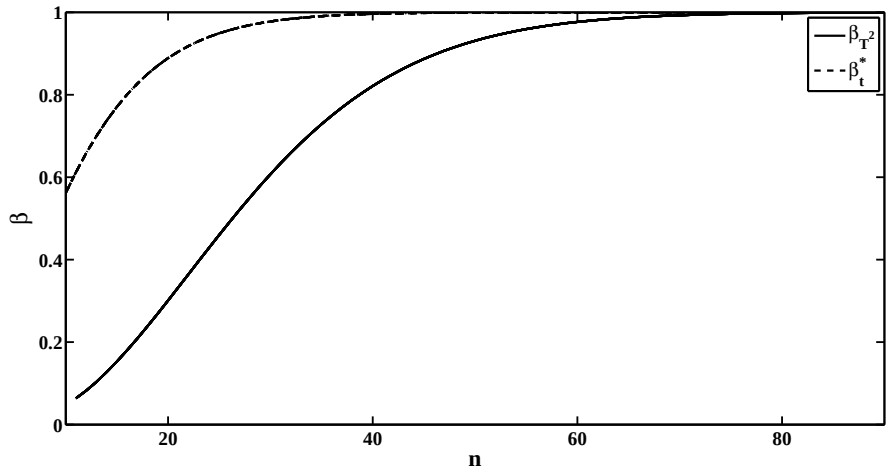


Figure 4.1: The power of T^2 and t^* tests versus the sample size n. The power of t^* is substantially larger than power of T^2 for small sample sizes, while for large sample sizes both power functions reach $\beta = 1$. Here $\alpha = 0.05$, $K = 10$.

4.4 The z^+ and t^+ tests

Since the optimal weighting vector $\boldsymbol{\omega}^*$ depends on the unknown model parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, suitable estimates must be derived to select the weighting vector. In this work, we propose estimating the weighting vector using prior information and data collected from a pilot study conducted prior to the main study. Here, we assume that the pilot and main study observations, $\boldsymbol{x}_p$ and $\boldsymbol{x}$, respectively, are sampled under the same conditions. That is, we let the $K-$dimensional pilot observations, $\boldsymbol{x}_{pi}$, be independent multivariate normal random variables

$$\boldsymbol{x}_{pi} \sim N_K\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right), \quad i = 1, 2, ..., n_p \qquad (n_p \geq 0), \tag{4.12}$$

with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

Prior information, $\mathcal{I}_0$, elicited from previous studies and expert clinical opinion is used to inform standard conjugate multivariate priors for the response mean

and covariance matrix. For the mean response we use the prior

$$(\boldsymbol{\mu} \mid \boldsymbol{\Sigma},\, \mathcal{I}_0) \sim N_K\left(\boldsymbol{m}_0,\, \boldsymbol{\Sigma}/n_0\right), \tag{4.13}$$

where $\boldsymbol{m}_0$ represents a prior estimate of the value of $\boldsymbol{\mu}$ and n_0 corresponds to the number of observations on which this prior estimate is based. When $\boldsymbol{\Sigma}$ is unknown, we use the inverse-Wishart $K \times K$ prior

$$(\boldsymbol{\Sigma} \mid \mathcal{I}_0) \sim IW_{K\times K}\left(\nu_0,\, \boldsymbol{S}_0^{-1}\right), \tag{4.14}$$

where ν_0 and $\boldsymbol{S}_0$ respectively represent the degrees of freedom and the (positive definite) scale matrix of the inverse-Wishart prior.

Under this standard Bayesian model (for example see [Gelman et al., 2004]), the posterior distribution of $\boldsymbol{\mu}$ given the information set, $\mathcal{I} = \{\mathcal{I}_0, \boldsymbol{X}_p\}$, and conditional on $\boldsymbol{\Sigma}$, is $K-$dimensional Gaussian with covariance matrix $\boldsymbol{\Sigma}/(n_0 + n_p)$ and mean

$$\boldsymbol{m} = \frac{n_0}{n_0 + n_p}\boldsymbol{m}_0 + \frac{n_p}{n_0 + n_p}\bar{\boldsymbol{x}}_p, \tag{4.15}$$

where $\bar{\boldsymbol{x}}_p = (n_p)^{-1}\sum_{i=1}^{n_p}\boldsymbol{x}_{pi}$ is the multivariate sample mean of the pilot data. The posterior distribution of the covariance matrix $(\boldsymbol{\Sigma}|\mathcal{I})$ is inverse-Wishart with $\nu = n_0 + n_p - 1$ degrees of freedom and scale matrix

$$\boldsymbol{S} = \boldsymbol{S}_0 + (n_p - 1)\boldsymbol{S}_{\boldsymbol{x}_p} + \frac{n_0 n_p}{n_0 + n_p}(\bar{\boldsymbol{x}}_p - \boldsymbol{m}_0)(\bar{\boldsymbol{x}}_p - \boldsymbol{m}_0)^T, \tag{4.16}$$

where $\boldsymbol{S}_{\boldsymbol{x}_p} = (n_p - 1)^{-1}\sum_{i=1}^{n_p}\left(\boldsymbol{x}_{pi} - \bar{\boldsymbol{x}}_p\right)\left(\boldsymbol{x}_{pi} - \bar{\boldsymbol{x}}_p\right)^T$ is the sample covariance matrix of the pilot data.

In particular, the modeling assumptions (4.12), (4.13) and (4.14) encompass the following special cases: (i) no prior information is available or $n_0, \nu_0 \to 0$, (ii) no pilot data is used to select the weighting vector or $n_p = 0$. In case (i) we require $n_p - 1 > K$ to obtain a positive definite $\boldsymbol{S}$. However, it should be stressed that these

methods rely on prior information for the effect direction and if no prior information is available the investigators may find it more appropriate to use the non-directional Hotelling's T^2 test.

We wish to use this information to select the weighting vectors $\boldsymbol{w}_j$ optimally. Optimality here is expressed in terms of predictive power of the test. The predictive power of the z and t test given the information set $\mathcal{I}$ is defined as

$$b_z = Pr\left(|z| > z_{\alpha/2} | \, \mathcal{I}\right), \tag{4.17}$$

$$b_t = Pr\left(|t| > t_{n-1,\alpha/2} | \, \mathcal{I}\right). \tag{4.18}$$

The following theorem provides the weighting vectors maximising the predictive power of the z and t tests given the information set $\mathcal{I}$.

Theorem 4.4.1. *Under* (4.1), (4.12) *and* (4.13), *the predictive power,* b_z, *of the* z *test in* (4.17) *is maximized with respect to the weighting vector* $\boldsymbol{w}$ *if and only if* $\boldsymbol{w}$ *is proportional to*

$$\boldsymbol{w}_{z^+} = \boldsymbol{\Sigma}^{-1}\boldsymbol{m}, \tag{4.19}$$

where $\boldsymbol{m}$ *as in* (4.15).

For $n_p \to \infty$, *the predictive power,* b_t, *of the* t *test in* (4.17) *is maximized with respect to the weighting vector* $\boldsymbol{w}$ *if and only if* $\boldsymbol{w}$ *is proportional to*

$$\boldsymbol{w}_{t^+} = \boldsymbol{S}^{-1}\boldsymbol{m}, \tag{4.20}$$

where $\boldsymbol{m}$, $\boldsymbol{S}$ *define as in* (4.15) *and* (4.16), *respectively.*

Proof. The z statistic can be written as

$$z = \theta\sqrt{n} + e, \qquad e \sim N(0,1),$$

where θ is defined in (4.8). Under (4.1), (4.13), $(\theta \mid \mathcal{I})$ is normally distributed with

variance $(n_0 + n_p)^{-1}$ and mean

$$\hat{\theta} = \boldsymbol{w}^T \boldsymbol{m}/\sigma.$$

Thus, $(z \mid \mathcal{I}) \sim N\left(\hat{\theta}\sqrt{n}, 1 + n/(n_0 + n_p)\right)$. The result is then proved using the same steps as in theorem 4.3.1 where θ is replaced by $\hat{\theta}$.

For the t test, we compute the asymptotic distribution of $t \mid \mathcal{I}$ as $n_p \to \infty$. By Bayes' rule, under (4.1), (4.13), (4.14), we have that,

$$(\boldsymbol{x}_i \mid \mathcal{I}) \sim t_K\left(\nu_0 + n_p - K + 1, \boldsymbol{m}, c\boldsymbol{S}\right), \qquad i = 1, 2, \ldots, n$$

where

$$c = \frac{n_0 + n_p + 1}{(n_0 + n_p)(\nu_0 + n_p - K + 1)}.$$

Hence for $n_p \to \infty$,

$$(\boldsymbol{x}_i \mid \mathcal{I}) \sim N_K\left(\boldsymbol{m}, \boldsymbol{S}\right), \qquad i = 1, 2, \ldots, n$$

Using the last result and corollary 7.2.3 in Anderson [1984] we have, for $\bar{y} = \boldsymbol{w}^T \bar{\boldsymbol{x}}$, that

$$(\bar{y}|\mathcal{I}) \sim N\left(\boldsymbol{w}^T \boldsymbol{m}, \boldsymbol{w}^T \boldsymbol{S} \boldsymbol{w}/n\right). \tag{4.21}$$

Using the same result we have that for $\boldsymbol{S}_{\boldsymbol{x}} = (n-1)^{-1} \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$,

$$(\boldsymbol{S}_{\boldsymbol{x}}|\mathcal{I}) \sim W_{K\times K}\left(n-1, \boldsymbol{S}/(n-1)\right)$$

and hence, for $s^2 = \boldsymbol{w}^T \boldsymbol{S}_{\boldsymbol{x}} \boldsymbol{w}$, by proposition 3.4.2 in Mardia et al. [1979] that

$$\left(\frac{s^2}{\boldsymbol{w}^T \boldsymbol{S} \boldsymbol{w}/(n-1)} | \mathcal{I}\right) \sim \chi^2_{n-1}, \text{ independent of } \bar{y}. \tag{4.22}$$

From (4.21) and (4.22), it follows that the asymptotic distribution of the t statistic

in (4.5) given $\mathcal{I}$ and for $n_p \to \infty$ can be written as

$$t \mid \mathcal{I} = \frac{z + \hat{\vartheta}\sqrt{n}}{\sqrt{x^2/(n-1)}},$$

where $z \sim N(0,1)$ and $x^2 \sim \chi^2_{n-1}$, that is, $t \mid \mathcal{I}$, for $n_p \to \infty$, is approximately non-central t distributed with non-centrality parameter

$$\hat{\vartheta}\sqrt{n} = \frac{\boldsymbol{w}^T\boldsymbol{m}}{\sqrt{\boldsymbol{w}^T\boldsymbol{S}\boldsymbol{w}}}\sqrt{n}$$

and $n-1$ degrees of freedom. The result follows by replacing θ with $\hat{\vartheta}$ in the proof of theorem 4.3.1.

□

In the weighting vectors $\boldsymbol{w}_{z^+}$ and $\boldsymbol{w}_{t^+}$, compared to $\boldsymbol{\omega}^*$, the unknown model parameters are replaced by their posterior estimates. Hence, outcomes which based on the elicited prior knowledge and the observed pilot data are expected to have larger effects, smaller variances and smaller correlations, they receive greater weights.

It is important to stress that the above information is collected and the weighting vectors are fixed prior to the start of the main study. The pilot data is only used for selecting the weighting vector and possibly for selecting other aspects of the design of the study (for example the sample size n). Herein we will refer to z^+ and t^+ as representing the linear combination z and t test statistics with weighting vectors $\boldsymbol{w}_{z^+}$ and $\boldsymbol{w}_{t^+}$, respectively. The fact that the weighting vectors $\boldsymbol{w}_{z^+}$ and $\boldsymbol{w}_{t^+}$ are chosen before the initiation of the main study implies that the z^+ and t^+ tests control the false positive rate at their nominal significance level. We next consider an alternative approach to our problem. Note that in the next section we use a slightly different notation than the rest of the thesis to suit better Bayesian methodology.

4.5 Bayesian multivariate tests

For purposes of comparison, it might be interesting to consider an alternative fully Bayesian approach. First, it is useful to make clear the reasons why in this thesis we focus on hybrid methods which use prior information to select design parameters such as the sample size or, in our problem, the weighting vector, but derive statistical inference using classical frequentist methodology.

The primary reason is that in our motivating applications, that is, neuroimaging but more generally biomedical studies and clinical trials, the control of the frequentist or long-run type I error is of paramount importance. In Bayesian methodology, frequentist or long-run probabilities are conceptually irrelevant and their computation is often complicated and dependant on prior specification. Furthermore, control of these probabilities is often not attained. Another reason for not following a fully Bayesian methodology is that, in our motivating examples, the use of prior distribution for statistical inference is often controversial and in clinical trials is typically prohibited. In contrast, the use of prior information for the design of the study, as in our methods, is widely acceptable. Finally, Bayesian tests and especially those performed using Bayes Factors tend to be complicated and unfamiliar to practitioners and the latter introduces further complexities in implementation.

However, in cases where these concerns are less important, one may wish to follow a Bayesian methodology as this allows a more natural way to synthesize information from various sources and to assess the data evidence against or in favour of the null hypotheses based on posterior probabilities.

There are several ways to derive a Bayesian test. First, one may use the posterior probabilities $Pr(H_0|\boldsymbol{x})$ and $Pr(H_1|\boldsymbol{x})$. This results, for $\boldsymbol{\Sigma}$ known, to rejecting H_0 for large values of the posterior Mahalanobis distance $(n_0 + n)\boldsymbol{m}'\boldsymbol{\Sigma}^{-1}\boldsymbol{m}$ where $\boldsymbol{m}$ the posterior mean of $\boldsymbol{\mu}$. As [Berger,1985] comments "posterior probabilities of hypotheses are the primary Bayesian measures in testing problems".

Bayesian tests are also often derived using Bayes Factors,

$$\begin{aligned} BF &= \frac{\text{posterior odds ratio}}{\text{prior odds ratio}} \\ &= \frac{Pr(H_0|\boldsymbol{x})/Pr(H_1|\boldsymbol{x})}{Pr(H_0)/Pr(H_1)}. \end{aligned}$$

expressing the odds for H_0 to H_1 given by the data (but in most cases also affected by the prior). The Bayes Factor can be re-written in the more convenient form,

$$BF_{01} = \frac{Pr(\boldsymbol{x}|H_0)}{Pr(\boldsymbol{x}|H_1)}$$

which shows the connections of BF with the likelihood ratio. In our setting,

$$BF_{01} = \frac{f(\boldsymbol{x};\boldsymbol{\mu}=\mathbf{0}_K)}{\int_{\boldsymbol{\mu}\neq 0} f(\boldsymbol{x};\boldsymbol{\mu})f(\boldsymbol{\mu}|\mathcal{I}_0)d\boldsymbol{\mu}} = \frac{L(\boldsymbol{\mu}=\mathbf{0}_K;\boldsymbol{x})}{\int L(\boldsymbol{\mu};\boldsymbol{x})f(\boldsymbol{\mu})d\boldsymbol{\mu}}$$

where $f(\boldsymbol{x};\boldsymbol{\mu})$ and $L(\boldsymbol{\mu};\boldsymbol{x})$ the density of the data $\boldsymbol{x}$ and the likelihood function of the parameter value $\boldsymbol{\mu}$ while $f(\boldsymbol{\mu})$ the prior density of $\boldsymbol{\mu}$[1]. We consider here the case of known Σ and we set the prior

$$f(\boldsymbol{\mu}) = \begin{cases} p_0, & \text{for } \boldsymbol{\mu}=\mathbf{0}_K, \\ (1-p_0)\phi_{\boldsymbol{m}_0,\boldsymbol{T}_0}(\boldsymbol{\mu}), & \text{for } \boldsymbol{\mu}\neq\mathbf{0}_K, \end{cases}$$

where $\phi_{m_0,T_0}(\cdot)$ the density of multivariate normal distribution with mean $\boldsymbol{m}_0$ and covariance matrix $\boldsymbol{T}_0$. The latter ensures a non-zero value for the prior (and posterior) probability of H_0. Under this prior, we compute

$$BF_{01} = \left|\frac{T_0}{T_1}\right|^{1/2} \exp\left[-\frac{n}{2}\left\{\bar{\boldsymbol{x}}^T\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}} - (\bar{\boldsymbol{x}}-\boldsymbol{m}_0)^T\left(\boldsymbol{\Sigma}^{-1}\boldsymbol{T}_1\boldsymbol{T}_0^{-1}\right)(\bar{\boldsymbol{x}}-\boldsymbol{m}_0)\right\}\right],$$

[1]The restriction $\boldsymbol{\mu}\neq 0$ is omitted in the last equation as the value of the integral is unchanged with the addition of a single point of a continuous function in the integration area.

where

$$\boldsymbol{T}_1 = \left(\boldsymbol{T}_0^{-1} + n\boldsymbol{\Sigma}^{-1}\right)^{-1}$$

the covariance matrix of the posterior distribution of $\boldsymbol{\mu}$. Note that if we take

$$\boldsymbol{T}_0 = \boldsymbol{\Sigma}/n_0$$

then the form of the Bayes Factor simplifies to

$$BF_{01} = \left(\frac{n_1}{n_0}\right)^{1/2} \exp\left[-\frac{n}{2}\left\{\bar{\boldsymbol{x}}^T\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}} - \frac{n_0}{n_1}\left(\bar{\boldsymbol{x}} - \boldsymbol{m}_0\right)^T\boldsymbol{\Sigma}^{-1}\left(\bar{\boldsymbol{x}} - \boldsymbol{m}_0\right)\right\}\right],$$

where $n_1 = n_0 + n$.

In order to get a better idea of how the Bayes Factor approach compares with our methodology we next provide a short empirical study. Here we fix the observation dimension $K = 5$, the pilot sample size $n_p = 8$, the main study sample size $n = 12$ (total sample size $n_T = n_p + n = 20$) and $T_0 = \Sigma/n_0$ with the prior sample size $n_0 = 5$. We approximate the power of z^+, β_{z^+}, and the power of the Bayes Factor test, β_{BF}, by the rate of rejections in $R = 20000$ iterations. For the Bayes Factor test, the null hypothesis H_0 is rejected for $BF_{10} = BF_{01}^{-1} > b = 8$. The latter is a very conservative boundary, since typically $b > 3$ is considered significant evidence against H_0 [Kass,1995].

In table 4.1 we provide the power of the above tests for various values of the model parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and the prior $\boldsymbol{m}_0$. The mean parameter $\boldsymbol{\mu}$ is taken under both H_0 ($\boldsymbol{\mu} = \mathbf{0}_K$) and H_1 ($\boldsymbol{\mu} = \mathbf{1}_K/\sqrt{K}$), while in both cases the priors $\boldsymbol{m}_0 = \mathbf{0}_K, \mathbf{1}_K/\sqrt{K}$ are applied. As we can see, the Bayes Factor test, even for the conservative boundary $b = 8$, can seriously inflate the type I error rate (see β_{BF} for $\boldsymbol{\mu} = \mathbf{0}_K$). The latter rate depends strongly on the prior specification and the covariance structure for the Bayes Factor test, in contrast to the z^+ test which controls type I error. Higher power is achieved by the Bayes Factor test (although

this must be considered in the light of a higher type I error rate).

Table 4.1: Power of z^+ and Bayes Factor test for various values of the model parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and the prior $\boldsymbol{m}_0$. Here, $K = 5$, $n_p = 8$, $n = 12$ ($n_T = n_p + n = 20$), $\boldsymbol{T}_0 = \boldsymbol{\Sigma}/n_0$, $n_0 = 5$, $b = 8$.

$\boldsymbol{\Sigma}$	$\boldsymbol{\mu}$	$\boldsymbol{m}_0$	β_{z^+}	β_{BF}
$\boldsymbol{I}_K$	$\boldsymbol{0}_K$	$\boldsymbol{0}_K$	0.05	0.21
		$\boldsymbol{1}_K/\sqrt{K}$	0.05	0.06
	$\boldsymbol{1}_K/\sqrt{K}$	$\boldsymbol{0}_K$	0.77	0.99
		$\boldsymbol{1}_K/\sqrt{K}$	0.88	0.99
block diagonal	$\boldsymbol{0}_K$	$\boldsymbol{0}_K$	0.05	0.20
with blocks $\boldsymbol{I}_2$, $\boldsymbol{I}_3/2$		$\boldsymbol{1}_K/\sqrt{K}$	0.05	0.03
	$\boldsymbol{1}_K/\sqrt{K}$	$\boldsymbol{0}_K$	0.95	0.99
		$\boldsymbol{1}_K/\sqrt{K}$	0.98	0.99

The case of unknown $\boldsymbol{\Sigma}$ is more complicated and it is not considered here.

4.6 Discussion

The proposed methodology attempts to provide solutions satisfying the three main targets of hypothesis testing as discussed in section 4.1. Firstly, we believe that connecting the weighting vector to the prior knowledge and pilot data using a formal framework improves the interpretability of the conclusions. The investigators along with the outcome of rejection or acceptance of the null hypothesis, can easily understand why an outcome is more or less heavily weighted to derive the conclusion.

Furthermore, the type I error of these tests is controlled by simply fixing the weighting vector prior to the beginning of the main study. It is important to

acknowledge that prior information is used only to select the weighting vector and not for the final data evaluation. This procedure can be readily extended to allow for further design parameters to be decided using the prior information and the pilot data without inflating the type I error. The power performance of the z^+ and t^+, including comparisons to alternative global tests, will be studied in chapter 8. In section 8.4, we also consider applications to the neuroimaging examples described in sections 2.3.2 and 2.4.2.

In situations where the sample size is limited, conducting a pilot study external to the main study, using a part of the available sample, has some crucial disadvantages. The main disadvantage of these external pilot studies is that they introduce a trade-off with often no satisfactory solution. On the one hand, performing a small pilot study allows for a sufficient sized sample to be used in the main study, but may not provide sufficient information to design the latter. On the other hand, relatively large pilot studies, reduce substantially the sample size for the main study and raise ethical concerns for not directly using a substantial part of the observations for the final conclusions [Bauer and Köhne, 1994; Jennison and Turnbull, 2000].

In chapter 6 we develop methodology to replace the external pilot study with an internal pilot study. This internal pilot study can be seen as the first stage of a two-stage or, generally, a multi-stage study. Here the pilot data is used for both testing and to select the weighting vector of the second stage. Adaptive designs, which are considered in the next chapter, allow for such testing procedures to be performed without inflating the type I error.

Chapter 5

Adaptive designs

5.1 Introduction

All the global tests considered in chapters 3 and 4 assume a simple standard design. Specifically, they assume that the multivariate outcome, $\boldsymbol{x}$, is randomly sampled from a study conducted under the classical single-stage design (SSD). This is the textbook approach [Armitage et al., 2002; Friedman et al., 2010; Matthews and Farewell, 2007; Peacock and Peacock, 2011] for the design of experiments inclusive of those performed in biomedical research (for example clinical trials, epidemiological studies). Under SSD, the investigators define all aspects of the design (for example target population, sample size, statistical methodology) at the planning stage. The study is then conducted, the full dataset is collected and conclusions are lastly derived following every step of the *a priori* fixed design. Any intervention to the design is strictly prohibited to avoid bias of the final results.

There are various concerns arising under SSD, particularly in biomedical research. These mainly ethical and economical issues necessitate the monitoring of data acquisition with the possibility of early termination and interim design modifications. Sequential designs [Armitage, 1954; Wald, 1945] and later on adaptive designs [Bauer and Köhne, 1994; Proschan and Hunsberger, 1995] were developed

to address these issues and allow for several actions to be taken at an interim point, without contaminating the validity of the conclusions.

In this chapter, we provide background on sequential and adaptive designs attempting to outline important issues and developments in the field. Our main focus is the methodology underpinning the adaptive linear combination tests developed in later chapters, but also the group sequential global tests used mainly for comparison. We begin with the concept of sequential and adapting designs and then we describe the methodology for performing group-sequential and adaptive testing procedures. Finally, we briefly summarize applications of these designs and discuss their potential and the challenges to be overcome.

5.2 Early stopping and design adaptation

We next discuss the concepts that give rise to early stopping and interim design adaptations. As we explain next, the latter motivate the development of sequential and adaptive designs.

5.2.1 Early stopping

For studies, such as those performed in agriculture, where typically all the results are made simultaneously available after a certain period of time, it is reasonable to analyse data and derive conclusions only after the full dataset is collected. However, in biomedical studies, as well as in industrial applications, where data is accumulated gradually over a period of time, it is more natural to monitor results as observed [Jennison and Turnbull, 2000; Whitehead, 1997].

Especially for clinical trials, not only it is natural, but several administrative, economic and ethical reasons necessitate data monitoring. In this setting, it is important to check that the study is being conducted as planned (for example eligibility criteria are satisfied), while it is also crucial to ensure the safety of the administrated

treatments. Furthermore, for ethical and economic reasons, it is important to take full advantage of the available human and monetary resources. Therefore, it is often advantageous to terminate the study earlier than planned, if this is suggested by interim results. Such early stopping allows the remaining resources to be allocated to another study. Furthermore, if the new treatment is proved unsafe or ineffective, early termination ensures that the remaining subjects can switch to a better treatment, while if the new treatment is proved safe and more effective, the time for the treatment to become publicly available is shortened [Jennison and Turnbull, 2000].

The latter ethical and economical issues were the main motivations for extending sequential design methodology initially developed for industrial applications [Wald, 1945] to the medical field [Armitage, 1954, 1958]. However, the use of early sequential designs in medical studies was initially restrained by their demand to analyze data and decide for early stopping continuously, that is, after every observation. Group-sequential design (GSD) methodology developed later by Pocock [1977] and O'Brien and Fleming [1979] enhanced application of sequential designs, particularly in clinical trials. In group-sequential designs, interim analyses are performed only after every time a number of observations (group) is collected. This makes them practically less demanding, while most of the benefits of continuous sequential designs are retained [Jennison and Turnbull, 2000].

Interim analysis and early stopping are also permitted under adaptive design methodology. However, as we explain next, the main motivation for the adaptive designs, developed after the work of Bauer and Köhne [1994] and Proschan and Hunsberger [1995], was to permit for interim design modifications.

5.2.2 Design modifications

As we briefly discussed above, in the standard fixed single-stage studies, the investigators rely entirely on the information available at the planning stage to select the experimental design. The latter defines how the study is to be conducted as well

as how to analyze the results and derive statistical inference. Regarding the study conduct, amongst other things, the investigators define the target population, the number of participating subjects (sample size), the rule for allocating subjects to different treatment groups and the measurements to be taken. Considering statistical inference, the primary and secondary hypotheses to be studied are defined as well as the corresponding measures (often called endpoints) and statistical methodology to be used for their evaluation. For the conclusions of the study to remain valid under the traditional SSD, the study and the subsequent data analysis needs to be conducted following in every step the initially specified design [Armitage et al., 2002; Friedman et al., 2010].

However, in some cases, the information available at the planning stage is not sufficiently precise and reliable to define all aspects of the design. This is a problem arising even when primary design parameters, such as the sample size of the study, are to be determined. For example, to compute the sample size of a clinical trial with normally distributed responses, investigators are typically required to provide an estimate of the response variance as well as the value of the treatment effect described by regulatory authorities as the "minimal effect which has clinical relevance" [ICH, 1998]. However, in some cases, these values cannot be precisely defined, even after the imperative careful planning. A similar problem is also confronted in linear combination tests, in cases where the effect direction is not precisely known at the planning stage. In such circumstances, investigators may be reluctant to design the whole study based on imprecise estimates, as this might be considered as unethical and can proved to be inefficient [Bauer and Köhne, 1994; Kirby and Chuang-Stein, 2010].

One approach to this problem is to first perform a small external pilot study. This pilot study can be used to obtain information for various aspects of the design of the main study. In the tests developed in chapter 4, we follow this approach for performing linear combination tests. However, as discussed in section 4.6, this

approach has obvious limitations, especially in cases where only a restricted sample size is available.

A more satisfying solution is provided by adaptive designs as well as various developments of the classical group-sequential designs [Denne and Jennison, 2000; Jennison and Turnbull, 2003; Wittes and Brittain, 1990]. This can potentially mitigate the ethical and economic issues arising with external pilot studies and provide type I error control under certain design modifications. To deal with the issue of insufficient information at the planning stage, these designs suggest performing an *internal* rather than an external pilot study. That is, to consider the pilot study as the first stage of a two-stage or, generally, a multi-stage study. The pilot or first-stage data are then used for deriving interim decisions (for example early stopping with rejection/acceptance of the null hypothesis), but they can also be used for re-assessing and possibly modifying various aspects of the design. In a multi-stage design, interim analysis and design modifications can be performed sequentially each time a number of observations are collected.

Such designs can be seen as a method of allowing for modification, at interim analyses, of the initially planned design. Various authors suggest interim design modifications, under the above framework, to deal with new or unexpected results becoming available during the study conduction [Chi et al., 1999; Proschan and Hunsberger, 1995]. Such results may originate from other studies, but it can also arise from the collected data. For example, Chi et al. [1999] describe a single-stage study where at an interim analysis the observed treatment effects were substantially lower than expected but still clinically significant. This suggested an increase in the initially planned sample size as the latter was most likely not sufficient to derive a statistically significant outcome. However, if such a sample size re-calculation was not initially planned, it can be controversial and as various authors have shown (for example Chi et al. [1999]; Proschan and Hunsberger [1995]), if ignored in the final analysis, may substantially inflate the type I error. On the other hand, if the

observed results are ignored and the study continues as planned, the resources of the study might be wasted without reaching a convincing outcome. In the example provided by Chi et al. [1999], the study progressed as originally planned and the final analysis led to a statistically insignificant outcome. The above design framework which permits for interim design modifications provides a solution to these issues.

Note that this design framework conceptually consists of two sequential parts [Brannath et al., 2007]. The design of the first part is fixed, while the design of the second part can be changed based on the information that becomes available at the interim analysis, that is, interim data and possibly external information. Each part may consist of a single-stage study, as in a single group of observations, but, it is not unusual to conduct more than a single stage in the second part.

As we mentioned earlier, several authors proposed methodologies for performing interim design modifications within the group-sequential design framework. These approaches are typically characterized by a pre-planned adaptivity, that is, adaptation rules are completely pre-specified at the planning stage. These pre-planned adaptation rules are restricted to ensure that the form of the test statistics remains as in SSD. For example, modifications are often required to be independent of effect estimates and based solely on nuisance parameter estimates. If design modifications are more flexible, often using the observed effects estimates, adaptive design methodology is required to ensure type I error control.

In the next section, we describe the methodology for performing testing procedures within group-sequential and adaptive designs. We mainly focus on the issues related to the methodology developed in later chapters. Special attention is given to type I error control while power analysis of group sequential and adaptive tests is conducted in chapter 8.

5.3 Group-sequential testing

In group-sequential designs (GSD), interim analyses are performed after each group of a (typically pre-specified) number of observations is collected. At these interim analyses, a test is performed to decide whether to stop the study and either accept or reject the (primary) null hypothesis, H_0, or continue to the next stage. In the standard GSD, the maximum number of analyses is pre-specified and if the final analysis is reached, the study is terminated and a test is performed to either accept or reject H_0. The tests in these interim and final analyses are typically performed using the same (often likelihood-ratio) statistics used in SSD applied to all the collected observations.

We next illustrate this methodology with the group-sequential global tests also considered in later chapters.

5.3.1 Group-sequential global tests

Suppose that we wish to evaluate the global null hypothesis of no treatment effect $H_0 : \boldsymbol{\mu} = \mathbf{0}_K$ using a group-sequential test. We use a GSD where up to J analyses are sequentially performed after collecting $n_{(1)}, n_{(2)}, \ldots, n_{(J)}$ observations, where $n_{(j)} = n_1 + n_2 + \cdots + n_j$, $j = 1, 2, \ldots, J$. At the $j-$th interim analysis, we compute the value of a pre-specified global test statistic and this is compared to suitable rejection and acceptance stage-wise critical values to decide whether to continue the study or stop the study and either accept or reject the global null hypothesis. The test statistics are based on all the collected observations up to the interim analysis, while the stage-wise critical values, as we explain in the next section, are derived to control the overall error rates of the test.

The $j-$th stage linear combination z and t statistics are

$$z_{(j)} = \frac{\bar{y}_{(j)}}{\sigma/\sqrt{n_{(j)}}}, \qquad t_{(j)} = \frac{\bar{y}_{(j)}}{s_{(j)}/\sqrt{n_{(j)}}}, \tag{5.1}$$

where σ is the (population) variance of the linear combination $y = \boldsymbol{w}^T\boldsymbol{x}$, $(\boldsymbol{w} \neq 0)$ and $\bar{y}_{(j)}$, $s_{(j)}$ are the sample mean and sample variance of the linear combinations collected up to the $j-$th interim analysis. Henceforth, we call the group-sequential tests with stage-wise statistics as in (5.1) linear combination z_{GS} and t_{GS} or simply z_{GS} and t_{GS} tests.

Furthermore, the group-sequential multivariate χ^2_{GS} and Hotelling's T^2_{GS} tests have stage-wise statistics

$$\chi^2_{(j)} = n_{(j)}\bar{\boldsymbol{x}}^T_{(j)}\boldsymbol{\Sigma}^{-1}\bar{\boldsymbol{x}}_{(j)}, \qquad T^2_{(j)} = n_{(j)}\bar{\boldsymbol{x}}^T_{(j)}\boldsymbol{S}^{-1}_{\boldsymbol{x}_{(j)}}\bar{\boldsymbol{x}}_{(j)}, \;\; j = 1, 2, \ldots, J \tag{5.2}$$

where $\bar{\boldsymbol{x}}_{(j)}$, $\boldsymbol{S}_{\boldsymbol{x}_{(j)}}$ are respectively the sample mean and sample covariance matrix of all the collected observations up to the $j-$th interim analysis.

Critical values

The repeated multiple testing performed in GSDs, if ignored, may produce substantial inflation of the type I error. This problem can be solved by considering the multivariate joint distribution of the interim test statistics [Jennison and Turnbull, 2000]. Various forms of stage-wise critical values, controlling the tests' error rates, are proposed in the literature. Pocock [1977] and O'Brien and Fleming [1979] provide the classical fixed critical values, while Lan and DeMets [1983] propose the error spending method in which critical values are derived, or the α rate is "spent", based on the observed information levels. Eales and Jennison [1992], Eales [1995] and Barber and Jennison [2002] provide methods for optimally deriving critical values to minimize the average sample size used in the study, while controlling the error rates of the test. In chapter 8, we use critical values of the class developed by Wang and Tsiatis [1987] which include as special cases those obtained by Pocock [1977] and O'Brien and Fleming [1979].

Calculating the group sequential critical values often requires numerical com-

putation and, for this, various computer software (for example R, SPLUS, SAS, FORTRAN) provide appropriate packages, such as the gsDesign implemented in R [Anderson, 2013].

5.4 Adaptive testing

As in group-sequential tests, adaptive testing procedures perform interim analysis to decide whether to terminate the study and reject or accept H_0 after every time a group of observations is collected, that is, a stage of the study is completed. However, adaptive tests are constructed to permit for interim adaptations and for this a suitable framework ensuring type I error control is developed. Adaptive testing procedures are often expressed in terms of combination functions which are used to combine the stage-wise test statistics at each interim analysis. As explained in the following sections, adaptive combination tests that adhere to the conditional invariance principle (CIP) attain type I error control. Alternatively, adaptive tests are expressed in terms of conditional error functions (CEF) in which case the conditional error principle is often used to prove type I error control [Bretz et al., 2009].

5.4.1 Combination tests

We next describe the adaptive combination tests and the conditional invariance principle which provide type I error control in these tests. For simplicity, we consider only two-stage designs, but the methodology can be easily extended to designs with more stages. In particular, the adaptive combination tests developed in later chapters use a more general $J-$stage ($J \geq 2$) representation.

Here, the design of the first-stage is fixed at the planning stage, while the second-stage design can be decided at the interim. After each stage, the observed data is used to compute the value of the stage-wise statistics, T_1 and T_2, and the corresponding $p-$values, p_1 and p_2. Note that the form of the first-stage statistic,

T_1, is fixed at the planning stage, while the form of the second stage statistic T_2 can be decided at the interim. At the planning stage, a combination function, $C(p_1, p_2)$, monotonically increasing in both arguments, is specified. This combination function defines how the $p-$values of the two-stages are combined in the final analysis. In addition, at the planning stage, we define the acceptance and rejection critical values, $\alpha_{1,0}$, $\alpha_{1,1}$ and $\alpha_{2,1}$, that satisfy the type I error equation of the test described below.

The study is conducted as follows. After collecting the observations of the first stage, the $p-$value p_1 is computed and if $p_1 \leq \alpha_{1,1}$ or $p_1 \geq \alpha_{1,0}$ the study is stopped with rejection or acceptance of H_0, respectively. If $p_1 \in (\alpha_{1,1}, \alpha_{1,0})$, the study continues to the second stage. The design of the second stage is decided and the second stage is conducted. After all observations are collected, the study terminates and the second-stage $p-$value is computed and combined with p_1 using $C(p_1, p_2)$. The null hypothesis H_0 is then rejected if $C(p_1, p_2) \leq \alpha_{2,1}$, otherwise accepted. The two-stage combination test can be formally described as follows:

$$\left.\begin{array}{rl} \text{At the interim analysis,} & \\ \quad \text{if } p_1 \leq \alpha_{1,1}\,, & \text{stop study and reject } H_0, \\ \quad \text{if } p_1 \geq \alpha_{1,0}\,, & \text{stop study and accept } H_0, \\ \quad \text{otherwise,} & \text{continue to stage 2.} \\ \text{At the final analysis,} & \\ \text{if } C(p_1, p_2) \leq \alpha_{2,1}, & \text{stop study and reject } H_0, \\ \quad \text{otherwise,} & \text{stop study and accept } H_0. \end{array}\right\} \tag{5.3}$$

The above combination test is said to satisfy the conditional invariance principle (CIP) if, despite the dependence of the second stage $p-$value, p_2, on the interim data, $\boldsymbol{X}_1$, the conditional null distribution of p_2 given the interim data, $p_2|\boldsymbol{X}_1$, equals a pre-specified null distribution. In other words, the CIP is satisfied, if the conditional null distribution of $p_2|\boldsymbol{X}_1$ is invariant to the design adaptation. Under this requirement, type I error of this combination test is controlled despite the

adaptations [Brannath et al., 2012; Bretz et al., 2009].

Example: Normally distributed stage-wise statistics

We now illustrate the application of CIP at a combination test with normally distributed stage-wise test statistics. This example applies to the adaptive tests developed in later chapters. First suppose that no adaptations are performed. Then, the test statistics T_1 and $\tilde{T}_2$, under H_0, are standard normally distributed. Thus, the corresponding $p-$values, p_1 and $\tilde{p}_2$, are independent and uniformly distributed on $(0,1)$, as in $p_1, \tilde{p}_2 \sim U(0,1)$ [George and Mudholkar, 1990]. Suppose now that an adaptation of the second stage is performed (for example sample size increased) based on the interim data $\boldsymbol{X}_1$. The second-stage statistic is now changed to T_2. If, despite the adaptation, the conditional null distribution of the modified $p-$value, $p_2|\boldsymbol{X}_1$, is still uniform, $p_2|\boldsymbol{X}_1 \sim U(0,1)$, the type I error of the combination test is controlled.

In this example, the type I error equation of the test in (5.3) is

$$\alpha = \alpha_{1,1} + \int_{\alpha_{1,1}}^{\alpha_{1,0}} Pr_0\left(C(p_1,p_2) \leq \alpha_{2,1}|p_1\right) \mathrm{d}p_1, \tag{5.4}$$

where the index 0 (as in $Pr_0(\cdot)$) indicates that the above probabilities are computed under the null hypothesis. If CIP is satisfied and the rejection and acceptance critical values are chosen at the planning stage to satisfy (5.4), the above equation holds even if design adaptations are performed at the interim analysis [Bauer and Köhne, 1994; Brannath et al., 2002; Bretz et al., 2009].

Note that, in this example, CIP implies stochastic independence between the $p-$values p_1 and p_2, under H_0. However, for type I error control, a weaker argument is sufficient. This is that the conditional distribution $p_2|\boldsymbol{X}_1$ is stochastically larger than the uniform distribution of p_2 or, formally,

$$Pr_0(p_2 \leq u \,|\, \boldsymbol{X}_1) \;\leq\; Pr_0(p_2 \leq u) \;=\; u, \;\; u \in (0,1). \tag{5.5}$$

The latter distribution of p_2 is also described as the conditionally larger than uniform or $p-$clud distribution [Brannath et al., 2002].

Combination function

Several combination functions are proposed in the literature. Bauer and Köhne [1994] suggest the use of Fisher's product combination function [Fisher, 1970], $C(p_1, p_2) = p_1 p_2$, while Lehmacher and Wassmer [1999] suggest the use of the weighted inverse normal combination function [Liptak, 1958],

$$C(p_1, p_2) = 1 - \Phi(c_1 \Phi^{-1}(1 - p_1) + c_2 \Phi^{-1}(1 - p_2)). \tag{5.6}$$

Here $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution and c_1, c_2 pre-specified weights ($c_1^2 + c_2^2 = 1$). The latter weights reflect the importance of each stage in the test. If the weights $c_1 = \sqrt{n_1/(n_1 + n_2)}$, $c_2 = \sqrt{n_2/(n_1 + n_2)}$ are used and no adaptations are performed, the combination function in (5.6) is equal to the classical group-sequential statistic for normally distributed stage-wise test statistics. The latter two combination functions are the most commonly used in the literature (see Bauer and Einfalt [2006]).

Other combination functions, such as the sum of $p-$values [Chang, 2007], as well as variations of Fisher's product [Brannath et al., 2002] and the inverse normal [Chi et al., 1999] are also suggested. These combination functions are also used in meta-analysis (see [Hedges and Olkin, 1985]). Finally, in a development of combination tests, Brannath et al. [2002] propose a recursive combination test which allows for the number of interim analyses, the critical values, and the sample sizes to be chosen adaptively during the experiment.

5.4.2 Conditional error approach

The adaptive linear combination tests developed in this thesis (see chapter 6) are based on the combination tests in the previous section. However, to obtain a more

complete picture of adaptive designs, we describe this alternative way of describing adaptive tests which is based on conditional error functions (CEFs). The idea was originally developed for two-stage designs by Proschan and Hunsberger [1995] and for multi-stage designs with a more general principle in Müller and Schäfer [2001]; Müller and Schäffer [2004]. We restrict this description to two-stage tests, but generalizations to studies with more stages is straightforward.

The conditional error function $A(\boldsymbol{X}_1)$ is the conditional probability, given the interim data $\boldsymbol{X}_1$, of rejecting the null hypothesis H_0, when H_0 is true, that is,

$$A(\boldsymbol{X}_1) = Pr_0(\text{ reject } H_0|\boldsymbol{X}_1), \tag{5.7}$$

where here the index 0 is used to denote that the above conditional probability is computed under H_0. The conditional error function is set to satisfy

$$E(A(\boldsymbol{X}_1)) = \int A(\boldsymbol{X}_1)\mathrm{d}F_0(\boldsymbol{X}_1) \leq \alpha, \tag{5.8}$$

where $F_0(\boldsymbol{X}_1)$ is the null cumulative distribution function of the first-stage data.

The tests defined within this framework reject H_0 if and only if the second stage $p-$value $p_2 \leq A(\boldsymbol{X}_1)$. Note that, under this test, the second stage does not need to be performed if for a given $\boldsymbol{X}_1$, $A(\boldsymbol{X}_1) = 1$ or $A(\boldsymbol{X}_1) = 0$ in which case H_0 is respectively rejected or accepted from the first stage.

To control type I error, the CEF is defined at the planning stage and satisfies (5.8). In addition, the conditional null distribution of $p_2|\boldsymbol{X}_1$ should be equal to or stochastically larger than the uniform distribution ($p-$clud). It is then straightforward to see that, even if interim design adaptations are performed, the type I error rate is

$$Pr_0(p_2 \leq A(\boldsymbol{X}_1)) = \int Pr_0(p_2 \leq A(\boldsymbol{X}_1) \,|\, \boldsymbol{X}_1)\, \mathrm{d}F_0(\boldsymbol{X}_1) \leq E(A(\boldsymbol{X}_1)) \leq \alpha, \tag{5.9}$$

that is, the type I error is controlled. Note that the equality in the above expression is implied by the Total Probability Law, while the inequalities by the $p-$clud property of $p_2|\boldsymbol{X}_1$ and the expression in (5.8).

Combination tests can be re-written in terms of the conditional error function [Bretz et al., 2009; Posch and Bauer, 1999]. For example, the combination test in (5.3), with $C(p_1, p_2) = p_1 p_2$, can be re-written in terms of the following CEF,

$$A(p_1) = \begin{cases} 1, & \text{if } p_1 \leq \alpha_{1,1} \\ \alpha_{2,1}/p_1, & \text{if } p_1 \in (\alpha_{1,1}, \alpha_{1,0}) \\ 0, & \text{if } p_1 \geq \alpha_{1,0}. \end{cases} \tag{5.10}$$

5.5 Types of design modifications

Sequential and adaptive designs are employed for many kinds of design modifications. One common application is sample size re-calculation. Sample size is a very important issue especially in medical studies, where ethical and economical concerns require the study to be conducted safely and efficiently with no human or monetary resources being wasted. Various authors discuss the subject and provide methods of re-assessing and modifying, if necessary, sample size at interim analysis within the framework of group-sequential [Denne and Jennison, 2000; Gould, 2001; Kieser and Friede, 2000; Stein, 1945; Wittes and Brittain, 1990] and adaptive designs [Chi et al., 1999; Chuong-Stein et al., 2006; Cui and Wu, 2010; Lehmacher and Wassmer, 1999; Mehta and Patel, 2006; Mehta and Pocock, 2011; Proschan, 2009; Proschan and Hunsberger, 1995].

Another area which has attracted considerable attention is the response-adaptive sample allocation. These designs are constructed to allocate subjects to the different treatment groups based on the observed responses rather than a fixed plan. Play-the-winner biased allocation rules are being developed [Hu and Rosenberger, 2003; Hu and Zhang, 2004; Wei and Durham, 1978] while Jennison and Turnbull

[2001] and Zhu and Hu [2010] provide methods for monitoring studies with response-adaptive allocation using sequential tests methodology.

Adaptive designs are also used for treatment and endpoint selection. That is, to identify the most effective among a number of different treatments or doses or to identify the most important among a number of different endpoints. These two problems are often described as a hypothesis selection or a multiple testing problem [Hommel, 2001; Koenig et al., 2008; Miller, 2010; Vandemeulebroecke et al., 2010]. In this setting, adaptive designs are often combined with the closure principle of Marcus et al. [1976]. A related attractive application of these methods is to combine phase II and phase III clinical trials in a single study [Bretz et al., 2006; Jennison and Turnbull, 2007; Kimani et al., 2009; Schmidli et al., 2007; Stallard and Todd, 2011]. Phase II studies are often performed to select the most effective and safe among a group of treatments or doses. After the best of these treatments is identified, phase III trials are performed to confirm or disprove the superiority of this treatment over an established treatment. In seamless phase II/III studies, the phase II study is treated as the fixed first stage of an adaptive design, with phase III study as the adaptive second stage. The latter designs have obvious practical advantages including speeding up the drug development process.

5.5.1 Stage-wise statistic adaptation

In situations where various options for the form of the test statistic exist, as for example in multiple testing, a test statistic adaptation seems appealing. In this, the accumulated data are used at each interim analysis to select the subsequent stage-wise statistics. Despite the fact that many authors advocate the potential for such adaptations (such as Bauer and Köhne [1994]; Bretz et al. [2009]), there are only a few papers on the subject [Kieser et al., 2002; Lang et al., 2000; Neuhäuser, 2001]. Lang et al. [2000] propose an adaptive two-stage combination test for dose-finding, evaluating the global null hypothesis against monotonically ordered alternatives. At

the first stage, the likelihood-ratio test under the assumption of independent and monotonically ordered outcomes is used, while at the second stage they construct a linear combination test with weights equal to the maximum likelihood estimates under the latter assumption.

Furthermore, Neuhäuser [2001] propose an adaptive two-stage test for simultaneous evaluation of location and scale hypothesis, while Kieser et al. [2002] propose a two-stage design where the first stage data are used to perform a bootstrap comparison of the power of a selection of test statistics, where the most powerful test is used at the second stage.

5.6 Discussion: potential and challenges

Since the work of Bauer and Köhne [1994] and Proschan and Hunsberger [1995], great interest in adaptive designs, especially within the context of clinical trials, has been expressed in various forms. As Pong and Chow [2010] discuss, this is addressed by special sessions organised in many professional conferences and meetings, the large number of publications in the field with special issues published in many journals (such as Biometrics (Vol. 62, No. 3), Statistics in Medicine (Vol. 25, No. 19), Journal of Biopharmaceutical Statistics (Vol. 15, No. 4 and Vol. 17, No. 6), Biometrical Journal (Vol. 48, No. 4), and Pharmaceutical Statistics (Vol. 5, No. 2)) and the establishment of working groups and publication of white papers by various associations (see for example Gallo et al. [2006]) and regulatory authorities for clinical trials [EMEA, 2007; FDA, 2010].

As is recognized in the latter papers, adaptive designs offer opportunities for more efficient use of available resources in conducting a study. At the same time, it is widely understood that several dangers exist if adaptive designs are used without caution. EMEA [2007]; Gallo et al. [2006] and FDA [2010], which discuss the use of adaptive designs particularly in confirmatory clinical trials, stress that

careful planning of the latter trials should not be disregarded when applying an adaptive design. They also stress that, in this setting, interim analysis and design modifications should be performed only if absolutely necessary since they introduce a risk of damaging the integrity of the study. They clarify that the reasons for allowing for interim adaptations as well as the adaptation rules should be fully described and justified at the planning stage.

Furthermore, the methodology used by adaptive designs to accommodate flexibility (for example combination tests) introduces various complexities regarding interpretability and efficiency. Burman and Sonesson [2006] give examples where adaptive design methodology used naively and without caution can result in outcomes which are hard to interpret. Jennison and Turnbull [2003, 2006] and Tsiatis and Mehta [2003] proved that adaptive designs can be substantially less efficient than similar group-sequential designs. Jennison and Turnbull attribute this loss of efficiency not only to the use of combination (rather than likelihood-ratio) tests, but also to the suboptimal adaptation rules and the overreliance on interim estimates.

On the other hand, various authors attempt to improve the integrity and interpretability of adaptive designs. Liu et al. [2002] and Brannath et al. [2012] derive results that attempt to provide validation of adaptive tests in general situations. Liu et al. [2002] prove type I error control for any adaptation rule for which the set of possible adaptations is countable and the rule of adaptation is measurable. Brannath et al. [2012] show that type I error control of any adaptive design can be proved, if the conditional distribution of the second-stage data, given the first-stage data, can be described in terms of a regression model. Posch et al. [2003] provide adaptive designs that prohibit outcomes that are hard to interpret.

My personal view is that adaptive designs can be useful, particularly to deal with some difficult situations arising in medical studies. However, it is clear that adaptive designs require extra caution and, in some cases, extra effort not only in implementation but also in planning. It should be made clear that the lack of

knowledge at the planning stage, which makes adaptive designs necessary, comes with a cost in efficiency. The more information gathered at the planning stage, the less flexibility will be necessary and therefore the less efficiency cost will be paid. Therefore, the same effort with non-adaptive designs, if not more, should be made in careful planning. Furthermore, this cost in efficiency should be carefully studied, possibly through extensive simulations, at the planning stage. Last but not least, I believe that more work should be made in order to derive optimal adaptation rules.

The methodology developed in chapter 6 provides an example where power-optimality, within adaptive designs, is targeted, while type I error control is maintained. Furthermore, methodology for studying the efficiency of our adaptive tests is carefully developed in chapter 7 and used to perform extensive power analysis in chapter 8. Finally, in our testing procedures, the predictive-power-optimal adaptation rules are specified at the planning stage. This allows the use of our methods in contexts where pre-specification of the adaptation rule is required, but also allows further flexibility to be used as appropriate in situations where this is permitted.

Chapter 6

Adaptive linear combination tests

6.1 Introduction

In chapter 4, we developed linear combination tests which use a preliminary sample to derive the tests' weighting vector. This preliminary sample is collected from an external pilot study conducted prior to the main study. As we discussed in the previous chapter, external pilot studies raise various ethical and economical concerns mainly associated with the non-optimal use of available resources, especially if the latter are limited and expensive. Adaptive designs provide a solution to these issues, by turning the pilot study to the first stage of a two- or, generally, multi-stage design, in which after every stage a test is performed using the obtained data.

In the following, we employ adaptive designs to allow for sequential modifications of the linear combination test statistics based on accumulated data. The potential of test statistic adaptation methods, as we discussed in the previous chapter, has been advocated by various authors, but only a few papers exist on the subject. Furthermore, as we mention in chapter 5, various adaptive multiple testing procedures are developed in the literature. However, these methods target the de-

tection of one or some few independent local effects and they become conservative and inefficient in settings, such as the typical neuroimaging studies, where strong dependencies and large number of outcomes are present.

In our methods, an adaptive testing procedure is employed, where linear combination tests are used as stage-wise statistics. The linear combination statistics are initially constructed with weighting vector derived from prior information and then sequentially updated based on the observed data at subsequent interim analyses. While our tests maintain the two prime targets of adaptive designs, namely flexibility and type I error control [Brannath et al., 2012], they also target power optimality. In particular, our procedures are derived to maximize the predictive power of the test at each interim analysis.

The chapter is organised as follows. We start by formulating the class of $J-$stage linear combination z and t tests and then derive power-optimal, with respect to power, tests in this class. We then proceed to derive the proposed adaptive linear combination tests which are proved: (i) to be optimal with respect to predictive power and (ii) to control the type I error. The chapter is closed with discussion of the main results and outstanding issues.

6.2 Formulation of $J-$stage linear combination tests

In the following, we formulate $J-$stage linear combination z and t tests and define their error rate functions. We assume that the $K-$dimensional observation vectors $\boldsymbol{x}_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijK})^T$ of subjects $i = 1, 2, \ldots, n_j$, participating in stage j, $j = 1, 2, \ldots, J$, of the study, are independent and identically distributed Gaussian random variables

$$\boldsymbol{x}_{ij} \sim N_K(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{6.1}$$

with mean $\boldsymbol{\mu}$ and covariance matrix the positive definite $\boldsymbol{\Sigma}$.

We wish to test the global null hypothesis of no treatment effect against the

two-sided alternative,

$$H_0 : \boldsymbol{\mu} = \mathbf{0}_K \quad \text{versus} \quad H_1 : \boldsymbol{\mu} \neq \mathbf{0}_K.$$

As in the single-stage design, the methods which follow equally apply to both the two-sample setting with common covariance matrix and the setting of paired multivariate observations. We continue with the one-sample presentation to simplify notation.

The observation vectors $\boldsymbol{x}_{ij}$, $i = 1, 2, \ldots, n_j$, of the subjects participating in the $j-$th stage are projected on the non-zero weighting vector $\boldsymbol{w}_j = (w_{j1}, \ldots, w_{jK})^T$ and the projection magnitudes form the linear combinations

$$y_{ij} = \boldsymbol{w}_j^T \boldsymbol{x}_{ij}, \qquad i = 1, 2, \ldots, n_j, \;\; j = 1, 2, \ldots, J. \tag{6.2}$$

Note that the linear combinations y_{ij} for subjects $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, J$ are independent Gaussian random variables

$$y_{ij} \sim N(\mu_j, \sigma_j^2), \tag{6.3}$$

where $\mu_j = \boldsymbol{w}_j^T \boldsymbol{\mu}$, $\sigma_j^2 = \boldsymbol{w}_j^T \boldsymbol{\Sigma} \boldsymbol{w}_j$. Under H_0, $y_{ij} \sim N(0, \sigma_j^2)$, $i = 1, 2, \ldots, n_j$, independent.

The stage-wise z and t statistics for testing H_0 against H_1 using the random sample of linear combinations y_{ij}, $i = 1, \ldots, n_j$, when $\boldsymbol{\Sigma}$ is either known or unknown are respectively

$$z_j = \frac{\bar{y}_j}{\sigma_j / \sqrt{n_j}}, \qquad t_j = \frac{\bar{y}_j}{s_j / \sqrt{n_j}} \, . \tag{6.4}$$

Here,

$$\bar{y}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}, \qquad s_j^2 = \frac{1}{n_j - 1} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_j)^2 \tag{6.5}$$

are the sample mean and sample variance of the linear combinations, respectively.

The two-sided stage-wise p values of the z and t tests are respectively $p_{z_j} = 2\Phi(-|z_j|)$ and $p_{t_j} = 2\Psi_{\nu_j}(-|t_j|)$.

Under assumption (6.1), the stage-wise z and t statistics, z_j, t_j, $j = 1, \ldots, J$ are respectively normally and non-centrally t distributed, $z_j \sim N(\bar{\theta}_j, 1)$ and $t_j \sim t_{n_j-1}(\bar{\theta}_j)$ with location parameter

$$\bar{\theta}_j = \theta_j \sqrt{n_j}, \qquad \theta_j = \frac{\mu_j}{\sigma_j} = \frac{\boldsymbol{w}_j^T \boldsymbol{\mu}}{\sqrt{\boldsymbol{w}_j^T \boldsymbol{\Sigma} \boldsymbol{w}_j}}. \tag{6.6}$$

Under H_0, the stage-wise z and t statistics are standard normal and Student's t random variables, that is, $z_j \sim N(0,1)$ and $t_j \sim t_{n_j-1}$ and the $p-$values p_{z_j} and p_{t_j} uniformly distributed in $(0,1)$, $p_{z_j}, p_{t_j} \sim U(0,1)$, $j = 1, 2, \ldots, J$.

We propose performing the $J-$stage z and t tests using adaptive combination tests, which allow for interim design adaptations. As we have described in the previous chapter, the combination function $C(p_1, p_2, \ldots, p_j)$, $j = 2, 3, \ldots, J$ is specified before the start of the study. Here, p_j is the $p-$value of the $j-$th stage, either p_{z_j} or p_{t_j}. In addition, the rejection and acceptance critical values $\alpha_{j,1}$, $j = 1, 2, \ldots, J$ and $\alpha_{j,0}$, $j = 1, 2, \ldots, J-1$ are also pre-selected to satisfy the type I error equation of the test described next.

After collecting the observations of the first stage, the $p-$value p_1 is computed and if $p_1 \leq \alpha_{1,1}$ or $p_1 \geq \alpha_{1,0}$ the study is stopped with rejection and acceptance of H_0, respectively. If $p_1 \in (\alpha_{1,1}, \alpha_{1,0})$, the study continues to the second stage. The design of the second stage is decided, the second stage is conducted and after collecting the second stage data, the second stage $p-$value, p_2, is combined with p_1 in $C(p_1, p_2)$. If $C(p_1, p_2) \leq \alpha_{2,1}$ or $C(p_1, p_2) \geq \alpha_{2,0}$ the study is terminated and H_0 is respectively rejected or accepted. If $C(p_1, p_2) \in (\alpha_{2,1}, \alpha_{2,0})$, the study continues to the next stage. The procedure continues in the same way until the study is terminated early or the final $J-$th stage is completed. In the latter case, the study is terminated with either rejection or acceptance of H_0 if respectively

$C(p_1, p_2, \ldots, p_J) \leq \alpha_{J,1}$ or $C(p_1, p_2, \ldots, p_J) > \alpha_{J,1}$. The $J-$stage test is formally described as follows:

$$\left.\begin{array}{lll} \text{At interim analysis } j = 1, 2, \ldots, J-1, & & \\ \quad \text{if } C(\boldsymbol{p}_j) \leq \alpha_{j,1}\,, & \text{stop study and reject } H_0, & \\ \quad \text{if } C(\boldsymbol{p}_j) \geq \alpha_{j,0}\,, & \text{stop study and accept } H_0, & \\ \quad \text{otherwise,} & \text{continue to stage j+1.} & \\ \text{At the final analysis J,} & & \\ \quad \text{if } C(\boldsymbol{p}_J) \leq \alpha_{J,1}, & \text{stop study and reject } H_0, & \\ \quad \text{otherwise,} & \text{stop study and accept } H_0. & \end{array}\right\} \quad (6.7)$$

Here, $\boldsymbol{p}_j = (p_1, p_2, \ldots, p_j)^T$ is the vector of $p-$values collected up to stage j, $j = 1, 2, \ldots, J$.

Several combination functions have been proposed in the literature (see section 5.4.1 on page 74). The formulation and results which follow use the Fisher's product function,

$$C(\boldsymbol{p}_j) = \prod_{l=1}^{j} p_l, \qquad j = 1, 2, \ldots, J. \quad (6.8)$$

However, our results equally apply to other combination functions including the inverse normal in (5.6) (page 77).

Herein we will refer to the $J-$stage tests with linear combination stage-wise z and t test statistics as the $J-$stage z and t tests, respectively. The power of the $J-$stage z or t test is $\beta = \sum_{j=1}^{J} \beta_j$ where,

$$\beta_1 = Pr\left(\, p_1 \leq \alpha_{1,1}\right), \quad \beta_j = Pr\left(C(\boldsymbol{p}_l) \in (\alpha_{l,1}, \alpha_{l,0})\ \forall\ l < j\ ;\ C(\boldsymbol{p}_j) \leq \alpha_{j,1}\right), \quad (6.9)$$

are respectively the first stage and $j-$th stage, $j = 2, 3, \ldots, J$, power functions (β, β_j are either β_z, β_{z_j} or β_t, β_{t_j}, respectively).

To control the type I error of the combination test, the rejection and accep-

tance critical values need to satisfy the type I error equation

$$\alpha = Pr_0\left(\, p_1 \leq \alpha_{1,1}\right) + \sum_{j=2}^{J} Pr_0\left(C(\boldsymbol{p}_l) \in (\alpha_{l,1}, \alpha_{l,0})\ \forall\, l < j\ ;\ C(\boldsymbol{p}_j) \leq \alpha_{j,1}\right), \quad (6.10)$$

where the index 0 (as in $Pr_0(\cdot)$) indicates that the above probabilities are computed under H_0. The type I error equation, for the Fisher's combination function, can be alternatively written as

$$\alpha = \alpha_{1,1} + \sum_{j=2}^{J} \int_{\alpha_{1,1}}^{\alpha_{1,0}} \int_{\alpha'_{2,1}}^{\alpha'_{2,0}} \cdots \int_{\alpha'_{j-1,1}}^{\alpha'_{j-1,0}} \alpha'_{j,1}\ \, \mathrm{d}p_{j-1} \ldots \mathrm{d}p_2 \mathrm{d}p_1, \quad (6.11)$$

where

$$\alpha'_{j,1} = \alpha_{j,1}/p_1 p_2 \cdots\ p_{j-1}, \qquad \alpha'_{j,0} = \alpha_{j,0}/p_1 p_2 \cdots\ p_{j-1} \quad (6.12)$$

the conditional rejection and acceptance boundaries, respectively, of stage j, $j = 2, 3, \ldots, J$. For the two-stage design the latter reduces to

$$\alpha = \alpha_{1,1} + \int_{\alpha_{1,1}}^{\alpha_{1,0}} \alpha_{2,1}/p_1\ \, \mathrm{d}p_1 = \alpha_{1,1} + \alpha_{2,1}\left(\log \alpha_{1,0} - \log \alpha_{1,1}\right). \quad (6.13)$$

Thus, for fixed size α of the test and first stage critical values $\alpha_{1,0}$, $\alpha_{1,1}$ the second stage critical value is

$$\alpha_{2,1} = \frac{\alpha - \alpha_{1,1}}{\log \alpha_{1,0} - \log \alpha_{1,1}}. \quad (6.14)$$

Here, $\alpha_{1,0} \in (\alpha, 1]$ and often substantially larger than α (for example $\alpha_{1,0} > 0.5$) to avoid inappropriately high probability of early acceptance. Further, $\alpha_{1,1}$ is chosen to be in $(0, \alpha)$. The same restrictions are applied for tests with more than two stages ($J > 2$).

6.3 Optimal $J-$stage linear combination tests

In this section, we develop a methodology for optimally deriving the weighting vectors of the stage-wise linear combinations in the $J-$stage adaptive test. Optimality,

here, is defined in terms of the power function of these tests as in (6.9).

For this we use a similar approach with the methodology used in chapter 4. In the latter, lemma 4.3.1 and 4.3.2 (see pp. 51, 52) were used to prove that the power of the single-stage z and t tests are increasing in the absolute value of the location parameter, θ, of the z and t statistics. We wish to extend this result to the $J-$stage adaptive linear combination z and t tests.

We start by proving the results for the two-stage tests. To reduce technical details to the minimum, we provide the proofs of the subsequent results regarding the t tests in Appendix A. However, it might be useful to note here that in these proofs, the results for the z test are generalised using the equation (4.10) (page 52) linking the non-central t distribution with the normal distribution.

The first step is to show that the power of the two-stage z and t test is increasing to the absolute value of θ_1. A difficulty in deriving this result arises from the fact that for some range of values of θ_1 an increase in $|\theta_1|$, results in increased chance to reject at the current stage, but also a decreased chance to continue to the next stage and thus smaller β_2. The required result is derived by showing that even for these range of values of $|\theta_1|$, the decrease (in absolute value) in β_2 is bounded above by the increase in β_1.

Lemma 6.3.1. *Under* (6.1)*, the power of the two-stage z and t test in* (6.7) *with combination function as in* (6.8) *is non-decreasing in the absolute value of θ_1 in* (6.6)*.*

Proof. For the first stage power we have that

$$\frac{\mathrm{d}\beta_{z_1}}{\mathrm{d}\theta_1} = \sqrt{n_1}\left(\phi\left(z_{\frac{\alpha_{1,1}}{2}} - \bar{\theta}_1\right) - \phi\left(-z_{\frac{\alpha_{1,1}}{2}} - \bar{\theta}_1\right)\right). \tag{6.15}$$

The second stage power is

$$\beta_{z_2} = \int_{\mathcal{C}_{z_1}} \beta_{c,z_2}(\theta_2; z_1)\phi_{\theta_1}(z_1)\mathrm{d}z_1, \tag{6.16}$$

where $\mathcal{C}_{z_1} = \left(-z_{\alpha_{1,1}/2}, -z_{\alpha_{1,0}/2}\right) \cup \left(z_{\alpha_{1,0}/2}, z_{\alpha_{1,1}/2}\right)$ the first stage continuation region and $\beta_{c,z_2}(\theta_2; z_1)$ the second stage conditional power

$$Pr\left(|z_2| \geq z_{\frac{\alpha'_{2,1}}{2}} \mid p_{z_1}\right) = 1 - \Phi\left(z_{\frac{\alpha'_{2,1}}{2}} - \bar{\theta}_2\right) + \Phi\left(-z_{\frac{\alpha'_{2,1}}{2}} - \bar{\theta}_2\right), \tag{6.17}$$

where $\alpha'_{2,1}$ as in (6.12) and $\bar{\theta}_2$ as in (6.6). The second stage power in (6.19) can be written as

$$\beta_{z_2} = \int_{\mathcal{C}^+_{z_1}} \beta_{c,z_2}(\theta_2; z_1) \left(\phi(z_1 - \bar{\theta}_1) + \phi(z_1 + \bar{\theta}_1)\right) \mathrm{d}z_1,$$

where $\mathcal{C}^+_{z_1} = \left(z_{\alpha_{1,0}/2}, z_{\alpha_{1,1}/2}\right)$. Thus,

$$\frac{\mathrm{d}\beta_{z_2}}{\mathrm{d}\theta_1} = \int_{\mathcal{C}^+_{z_1}} \beta_{c,z_2}(\theta_2; z_1) \frac{\mathrm{d}\left(\phi(z_1 - \bar{\theta}_1) + \phi(z_1 + \bar{\theta}_1)\right)}{\mathrm{d}\theta_1} \, \mathrm{d}z_1$$

and using that $\phi(z_1 - \bar{\theta}_1) = \phi(z_1 + \bar{\theta}_1) e^{2\bar{\theta}_1 z_1}$,

$$\begin{aligned}\frac{\mathrm{d}\left(\phi(z_1 - \bar{\theta}_1) + \phi(z_1 + \bar{\theta}_1)\right)}{\mathrm{d}\theta_1} &= \sqrt{n_1}\left((z_1 - \bar{\theta}_1)\phi(z_1 - \bar{\theta}_1) - (z_1 + \bar{\theta}_1)\phi(z_1 + \bar{\theta}_1)\right) \\ &= \sqrt{n_1}(e^{2\bar{\theta}_1 z_1} + 1)\phi(z_1 + \bar{\theta}_1)(z_1 \tanh(\bar{\theta}_1 z_1) - \bar{\theta}_1), \end{aligned} \tag{6.18}$$

where $\tanh(\bar{\theta}_1 z_1) = \frac{e^{2\bar{\theta}_1 z_1} - 1}{e^{2\bar{\theta}_1 z_1} + 1}$ is the hyperbolic tangent at $\bar{\theta}_1 z_1$. First note that

$$\frac{\mathrm{d}\tanh\left(\bar{\theta}_1 z_1\right)}{\mathrm{d}z_1} = \frac{4\bar{\theta}_1 e^{2\bar{\theta}_1 z_1}}{\left(e^{2\bar{\theta}_1 z_1} + 1\right)^2},$$

which is non-negative for $\theta_1 \geq 0$ and negative for $\theta_1 < 0$. Hence,

$$\frac{\mathrm{d}\left(z_1 \tanh \bar{\theta}_1 z_1 - \bar{\theta}_1\right)}{\mathrm{d}z_1}$$

is non-negative for $\theta_1 \geq 0$ and negative for $\theta_1 < 0$ ($z_1 \in \mathcal{C}^+_{z_1}$). Thus, for $\theta_1 \geq 0$, we

can identify the following three cases:

$$
\begin{aligned}
&(1) \qquad z_1 \tanh(\bar{\theta}_1 z_1) - \bar{\theta}_1 \geq 0, \ \ \forall\, z_1 \in \mathcal{C}_{z_1}^{+}, \\
&(2) \qquad z_1 \tanh(\bar{\theta}_1 z_1) - \bar{\theta}_1 \leq 0, \ \ \forall\, z_1 \in \mathcal{C}_{z_1}^{+}, \\
&(3) \qquad \exists\, \dot{z}_1 \in \mathcal{C}_{z_1}^{+} : z_1 \tanh(\bar{\theta}_1 z_1) - \bar{\theta}_1 \leq 0, \ \forall\, z_1 \in (z_{\frac{\alpha_{1,0}}{2}}, \dot{z}_1), \\
&\qquad\qquad\qquad\qquad z_1 \tanh(\bar{\theta}_1 z_1) - \bar{\theta}_1 > 0, \ \forall\, z_1 \in (\dot{z}_1, z_{\frac{\alpha_{1,1}}{2}})
\end{aligned}
$$

Case (1), by (6.18), implies that

$$
\frac{\mathrm{d}\phi(z_1 - \bar{\theta}_1) + \phi(z_1 + \bar{\theta}_1)}{\mathrm{d}\theta_1} \geq 0, \ \ \forall\, z_1 \in \mathcal{C}_{z_1}^{+}.
$$

This in turn implies that β_{z_2} and hence β_z is non-decreasing in θ_1 for $\theta_1 \geq 0$.

Case (2), by (6.18), implies that

$$
\frac{\mathrm{d}\phi(z_1 - \bar{\theta}_1) + \phi(z_1 + \bar{\theta}_1)}{\mathrm{d}\theta_1} \leq 0, \ \ \forall\, z_1 \in \mathcal{C}_{z_1}^{+}.
$$

This in turn implies that

$$
\frac{\mathrm{d}\beta_{z_2}}{\mathrm{d}\theta_1} \geq \beta_{c,z_2}(\theta_2; z_{\frac{\alpha_{1,1}}{2}}) \int_{z_{\frac{\alpha_{1,0}}{2}}}^{z_{\frac{\alpha_{1,1}}{2}}} \frac{\mathrm{d}\phi(z_1 - \bar{\theta}_1) + \phi(z_1 + \bar{\theta}_1)}{\mathrm{d}\theta_1}\, \mathrm{d}z_1.
$$

Thus, using (6.15), we have that

$$
\begin{aligned}
\frac{\mathrm{d}\beta_z}{\mathrm{d}\theta_1} \geq & \sqrt{n_1} \left(\left(1 - \beta_{c,z_2} \left(\theta_2; z_{\frac{\alpha_{1,1}}{2}} \right) \right) \left(\phi(z_{\frac{\alpha_{1,1}}{2}} - \bar{\theta}_1) - \phi(z_{\frac{\alpha_{1,1}}{2}} + \bar{\theta}_1) \right) \right. \\
& \left. + \beta_{c,z_2} \left(\theta_2; z_{\frac{\alpha_{1,1}}{2}} \right) \left(\phi(z_{\frac{\alpha_{1,0}}{2}} - \bar{\theta}_1) - \phi(z_{\frac{\alpha_{1,0}}{2}} + \bar{\theta}_1) \right) \right) \geq 0.
\end{aligned}
$$

Case (3), by (6.18), implies that

$$\frac{\mathrm{d}\left(\phi(z_1-\bar{\theta}_1)+\phi(z_1+\bar{\theta}_1)\right)}{\mathrm{d}\theta_1} \leq 0 \qquad \forall\, z_1 \in \left(z_{\frac{\alpha_{1,0}}{2}}, \dot{z}_1\right),$$
$$\frac{\mathrm{d}\left(\phi(z_1-\bar{\theta}_1)+\phi(z_1+\bar{\theta}_1)\right)}{\mathrm{d}\theta_1} > 0 \qquad \forall\, z_1 \in \left(\dot{z}_1, z_{\frac{\alpha_{1,1}}{2}}\right).$$

This in turn implies that

$$\begin{aligned}\frac{\mathrm{d}\beta_{z_2}}{\mathrm{d}\theta_1} \geq & \beta_{c,z_2}(\theta_2;\dot{z}_1)\int_{z_{\frac{\alpha_{1,0}}{2}}}^{\dot{z}_1} \frac{\mathrm{d}\left(\phi(z_1-\bar{\theta}_1)+\phi(z_1+\bar{\theta}_1)\right)}{\mathrm{d}\theta_1}\mathrm{d}z_1 \\ & + \beta_{c,z_2}(\theta_2;\dot{z}_1)\int_{\dot{z}_1}^{z_{\frac{\alpha_{1,1}}{2}}} \frac{\mathrm{d}\left(\phi(z_1-\bar{\theta}_1)+\phi(z_1+\bar{\theta}_1)\right)}{\mathrm{d}\theta_1}\mathrm{d}z_1.\end{aligned}$$

Thus, using (6.15) we can identify the non-negative lower boundary

$$\begin{aligned}\frac{\mathrm{d}\beta_z}{\mathrm{d}\theta_1} \geq & \sqrt{n_1}\left(\left(1-\beta_{c,z_2}\left(\theta_2;\dot{z}_1\right)\right)\left(\phi(z_{\frac{\alpha_{1,1}}{2}}-\bar{\theta}_1)-\phi(z_{\frac{\alpha_{1,1}}{2}}+\bar{\theta}_1)\right)\right. \\ & \left. + \beta_{c,z_2}\left(\theta_2;\dot{z}_1\right)\left(\phi(z_{\frac{\alpha_{1,0}}{2}}-\bar{\theta}_1)-\phi(z_{\frac{\alpha_{1,0}}{2}}+\bar{\theta}_1)\right)\right) \geq 0.\end{aligned}$$

In all three possible cases, the power function of the two-stage z test, β_z, is non-decreasing in θ_1 for $\theta_1 \geq 0$. In the same way, we can show that β_z is decreasing in θ_1 for $\theta_1 < 0$ and the result follows. The result for the t test is provided in Appendix A. □

The second step is to show that power is increasing in the second stage location parameter.

Lemma 6.3.2. *Under* (6.1)*, the power of the two-stage z and t tests in* (6.7) *with combination function as in* (6.8) *are non-decreasing in the absolute value of θ_2 in* (6.6).

Proof. The second stage power of the two-stage z test is

$$\beta_{z_2} = \int_{\mathcal{C}_{z_1}} \beta_{c,z_2}(\theta_2;z_1)\phi_{\theta_1}(z_1)\mathrm{d}z_1, \tag{6.19}$$

where β_{c,z_2} the second stage conditional power as in (6.17). By similar steps as in the proof of lemma 4.3.1, we can prove that the derivative of β_{c,z_2} with respect to θ_2 is non-negative for $\theta_2 \geq 0$ and negative for $\theta_2 < 0$, and the result follows. The result for the t test is provided in Appendix A. □

The next lemma extends the latter results to the the $J-$stage z and t tests. The proof is based on re-writing the stage-wise power functions in a similar form to the stage-wise power functions of two-stage tests and then use recursively lemmas 6.3.1 and 6.3.2.

Lemma 6.3.3. *Under* (6.1)*, the power of the* $J-$*stage* z *and* t *tests in* (6.7) *with combination function as in* (6.8) *is non-decreasing in the absolute value of* θ_j *in* (6.6)*,* $j = 1, 2, \ldots, J$*.*

Proof. The $j-$th stage power of the z test can be written as

$$\beta_{z_j} = \int_{\mathcal{C}_{\boldsymbol{z}_{(j-1)}}} \beta_{c,z_j}(\theta_j; \boldsymbol{z}_{(j-1)}) f_{\boldsymbol{\theta}_{(j-1)}}(\boldsymbol{z}_{(j-1)})\, \mathrm{d}\boldsymbol{z}_{(j-1)},$$

where $f_{\boldsymbol{\theta}_{(j)}}(\boldsymbol{z}_{(j)})$ is the joint density of the vector of the stage-wise $z-$statistics up to stage j, $\boldsymbol{z}_{(j)} = (z_1, z_2, \ldots, z_j)$, with parameters $\boldsymbol{\theta}_{(j)} = (\theta_1, \theta_2, \ldots, \theta_j)$. Here, $\mathcal{C}_{\boldsymbol{z}_{(j)}} = \mathcal{C}_{z_1} \cap \mathcal{C}_{z_2} \cap \cdots \cap \mathcal{C}_{z_j}$ with $\mathcal{C}_{z_l} = (-z_{\alpha'_{l,1}/2}, -z_{\alpha'_{l,0}/2}) \cup (z_{\alpha'_{l,0}/2}, z_{\alpha'_{l,1}/2})$ the $l-$th stage continuation region, $\alpha'_{l,1}$, $\alpha'_{l,0}$, as in (6.12), $l = 1, 2, \ldots, j$, $j = 2, 3, \ldots, J$ ($\alpha'_{1,1} = \alpha_{1,1}$, $\alpha'_{1,0} = \alpha_{1,0}$).

We can prove that the power of the $J-$stage test is non-decreasing in $|\theta_J|$ by considering the derivative of $\beta_{c,z_J}(\theta_J; \boldsymbol{z}_{(J-1)})$ as in proof of lemma 6.3.2. Considering θ_{J-1}, the power of the last two stages, $\beta_z^{(J-1)} = \beta_{z_{J-1}} + \beta_{z_J}$, can be written as

$$\beta_z^{(J-1)} = \int_{\mathcal{C}_{\boldsymbol{z}_{(j-2)}}} \beta_{c,z}^{(J-1)}(\boldsymbol{\theta}^{(J-1)}; \boldsymbol{z}_{(J-2)}) f_{\boldsymbol{\theta}_{(J-2)}}(\boldsymbol{z}_{(J-2)})\, \mathrm{d}\boldsymbol{z}_{(J-2)},$$

where $\boldsymbol{\theta}^{(J-l)} = (\theta_{J-l}, \theta_{J-l+1}, \ldots, \theta_J)$, $l = 1, 2, \ldots, J-1$, and

$$\beta_{c,z}^{(J-1)}(\boldsymbol{\theta}^{(J-1)}; \boldsymbol{z}_{(J-2)}) = \beta_{c,z_{J-1}}(\theta_{J-1}; \boldsymbol{z}_{(J-2)}) + \int_{\mathcal{C}_{z_{J-1}}} \beta_{c,z_J}(\theta_J; z_{J-1}) f_{\theta_{J-1}}(z_{J-1}) \mathrm{d}z_{J-1}, \tag{6.20}$$

the conditional power of the last two stages given the value of the test statistics of the previous stages, $\boldsymbol{z}_{(J-2)}$. Therefore, using the same steps as in lemma 6.3.1, with the power of the two-stage z test replaced by $\beta_{c,z}^{(J-1)}(\boldsymbol{\theta}^{(J-1)}; \boldsymbol{z}_{(J-2)})$ in (6.20), it can be shown that the power of the $J-$stage test is increasing in $|\theta_{J-1}|$.

Considering θ_{J-2}, the power of the last three stages, $\beta_z^{(J-2)}$, is

$$\beta_z^{(J-2)} = \int_{\mathcal{C}_{\boldsymbol{z}_{(J-3)}}} \beta_{c,z}^{(J-2)}(\boldsymbol{\theta}^{(J-2)}; \boldsymbol{z}_{(J-3)}) f_{\boldsymbol{\theta}_{(J-3)}}(\boldsymbol{z}_{(J-3)}) \, \mathrm{d}\boldsymbol{z}_{(J-3)},$$

where

$$\beta_{c,z}^{(J-2)}(\boldsymbol{\theta}^{(J-2)}; \boldsymbol{z}_{(J-3)}) = \beta_{c,z_{J-2}}(\theta_{J-2}; \boldsymbol{z}_{J-3}) + \int_{\mathcal{C}_{z_{J-2}}} \beta_{c,z}^{(J-1)}(\boldsymbol{\theta}^{(J-1)}; \boldsymbol{z}_{(J-2)}) f_{\theta_{J-2}}(\boldsymbol{z}_{(J-2)}) \, \mathrm{d}\boldsymbol{z}_{(J-2)} \tag{6.21}$$

and $\beta_{c,z}^{(J-1)}(\boldsymbol{\theta}^{(J-1)}; \boldsymbol{z}_{(J-2)})$ as in (6.20). Thus, using the same steps as in lemma 6.3.1 where power of the two-stage test is replaced by $\beta_{c,z}^{(J-2)}$ in (6.21) we have that the power of the $J-$stage test is increasing in $|\theta_{J-2}|$.

Proceeding sequentially we can easily see that the power of the last l stages, $\beta_z^{(J-l)}$, is

$$\beta_z^{(J-l)} = \int_{\mathcal{C}_{\boldsymbol{z}_{J-l-1}}} \beta_{c,z}^{(J-l)}(\boldsymbol{\theta}^{(J-l)}; \boldsymbol{z}_{(J-l-1)}) f_{\boldsymbol{\theta}_{(J-l-1)}}(\boldsymbol{z}_{(J-l-1)}) \, \mathrm{d}\boldsymbol{z}_{(J-l-1)},$$

where,

$$\beta_{c,z}^{(J-l)}(\boldsymbol{\theta}^{(J-l)};\boldsymbol{z}_{(J-l-1)}) = \beta_{c,z_{J-l}}(\theta_{J-l};\boldsymbol{z}_{(J-l-1)}) + \int_{\mathcal{C}_{z_{J-l}}} \beta_{c,z}^{(J-l+1)}(\boldsymbol{\theta}^{(J-l+1)};\boldsymbol{z}_{(J-l)}) f_{\theta_{J-l}}(z_{J-l})\,\mathrm{d}z_{J-l}$$

is non-decreasing in $|\theta_{J-l}|$, $l = 1, 2, \ldots, J-1$.

Regarding the t test, the result follows using the same steps as above but with the Gaussian distribution replaced by the non-central t distribution and based on lemmas 6.3.1 and 6.3.2, respectively. □

Note that it can be straightforwardly shown that the above results hold for both one-sided stage-wise tests and for the inverse normal combination function. Note also that, similarly to the proof for the single-stage test, the result can be generalized to more general adaptive design settings.

Furthermore, as in the single-stage tests, lemma 6.3.3 implies that in order to maximize the power of the $J-$stage linear combination z and t tests with respect to the weighting vectors $\boldsymbol{w}_j$, it is sufficient to maximize the value of θ_j, $j = 1, 2, \ldots, J$. Using this result, we next derive the power-optimal weighting vector.

Theorem 6.3.1. *Under* (6.1)*, the power of the* $J-$*stage* z *and* t *tests in* (6.7) *with combination function as in* (6.8) *are maximized with respect to the weighting vectors* $\boldsymbol{w}_j$*,* $j = 1, 2, \ldots, J$*, if and only if the latter are proportional to* $\boldsymbol{\omega}^* = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}$*.*

Proof. The proof follows using the same steps as in the proof of theorem 4.3.1 (page 53) with θ replaced by θ_j and $\boldsymbol{w}$ replaced by $\boldsymbol{w}_j$, $j = 1, 2, \ldots, J$. □

The last result provides the optimal, in terms of power, weighting vector for the $J-$stage linear combination tests $\boldsymbol{\omega}^*$. The optimal weighting vector $\boldsymbol{\omega}^*$ expresses the multivariate treatment effect standardized with respect to the covariance matrix $\boldsymbol{\Sigma}$. The $J-$stage linear combination test with weighting vector $\boldsymbol{\omega}^*$ attain the maximum power for these tests and this result is used in chapter 8.

However, the optimal vector $\boldsymbol{\omega}^*$ depends on the unknown parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and therefore is also unknown. For selecting the weighting vectors, in practice, we propose using the information for $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, available at each interim analysis. The source of this information is the data collected from the stages completed before each interim analysis, but also prior information extracted from previous studies and expert clinical opinion. Predictive power allows the incorporation of this information into our procedures in a natural and plausible way. The weighting vectors are selected to maximise predictive power, hence they are optimal in terms of the current knowledge at each interim analysis.

Importantly, as we also explain in the next section, if equation (6.10) is satisfied, the type I error rate of these tests is controlled.

6.4 The adaptive z^+_{AD} and t^+_{AD} tests

As in z^+ and t^+ tests, we use the standard Gaussian-inverse-Wishart priors for the observation mean and covariance matrix,

$$(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \mathcal{I}_0) \sim N_K\left(\boldsymbol{m}_0, \boldsymbol{\Sigma}/n_0\right), \qquad (\boldsymbol{\Sigma} \mid \mathcal{I}_0) \sim IW_{K\times K}\left(\nu_0, \boldsymbol{S}_0^{-1}\right), \tag{6.22}$$

where $\boldsymbol{m}_0$ represents a prior estimate of the value of $\boldsymbol{\mu}$ and n_0 corresponds to the number of observations on which this prior estimate is based on, while ν_0 and $\boldsymbol{S}_0$ respectively represent the degrees of freedom and the scale matrix of the inverse-Wishart prior.

Under this standard Bayesian model (see Gelman et al. [2004]), the posterior distribution of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ given the information set $\mathcal{I}_j = \{\mathcal{I}_0, \boldsymbol{X}_{(j)}\}$, consisting of the prior information $\mathcal{I}_0$ and the data collected up to the $j-$th interim analysis $\boldsymbol{X}_{(j)} = [\boldsymbol{X}_1 \boldsymbol{X}_2 \ldots \boldsymbol{X}_j]$ is

$$(\boldsymbol{\mu} \mid \boldsymbol{\Sigma}, \mathcal{I}_j) \sim N_K\left(\boldsymbol{m}_j, \boldsymbol{\Sigma}/n_{(j)}\right), \qquad (\boldsymbol{\Sigma} \mid \mathcal{I}_j) \sim IW_{K\times K}\left(\nu_j, \boldsymbol{S}_j^{-1}\right). \tag{6.23}$$

Here,

$$\boldsymbol{m}_j = \frac{n_0\boldsymbol{m}_0 + n_{(j)}\bar{\boldsymbol{x}}_{(j)}}{n_0 + n_{(j)}}, \tag{6.24}$$

and

$$\boldsymbol{S}_j = \boldsymbol{S}_0 + \nu_{(j)}\boldsymbol{S}_{\boldsymbol{x}_{(j)}} + \frac{n_0 n_{(j)}}{n_0 + n_{(j)}}(\bar{\boldsymbol{x}}_{(j)} - \boldsymbol{m}_0)(\bar{\boldsymbol{x}}_{(j)} - \boldsymbol{m}_0)^T, \tag{6.25}$$

where $\nu_{(j)} = n_{(j)} - 1$, $\nu_j = \nu_0 + n_{(j)}$ with $n_{(j)} = n_1 + n_2 + \cdots + n_j$ and

$$\bar{\boldsymbol{x}}_{(j)} = \sum_{l=1}^{j}\sum_{i=1}^{n_l}\boldsymbol{x}_{il}/n_{(j)}, \qquad \boldsymbol{S}_{\boldsymbol{x}_{(j)}} = \frac{1}{n_j - 1}\sum_{l=1}^{j}\sum_{i=1}^{n_l}\left(\boldsymbol{x}_{il} - \bar{\boldsymbol{x}}_{(j)}\right)\left(\boldsymbol{x}_{il} - \bar{\boldsymbol{x}}_{(j)}\right)^T \tag{6.26}$$

are the sample mean and sample covariance matrix of $\boldsymbol{X}_{(j)}$. Note that, due to the positive definiteness of the prior estimates $\boldsymbol{S}_0$, the posterior estimates $\boldsymbol{S}_j$, $j = 1, 2, \ldots, J$, are also positive definite.

We wish to use this information to select the weighting vectors $\boldsymbol{w}_j$ optimally. Optimality here is expressed in terms of predictive power of the test. The predictive power for the first stage given the prior information set $\mathcal{I}_0$ is $b_1 = Pr\left(p_1 \leq \alpha_{1,1} \mid \mathcal{I}_0\right)$ and for the $j-$th stage, $j = 2, 3, \ldots, J$, given the information set $\mathcal{I}_{j-1}$ is

$$b_j = \begin{cases} 1\,, & \mathcal{I}_{j-1} \text{ such that } C(\boldsymbol{p}_l) \leq \alpha_{l,1} \text{ for } l \in \{1, 2, \ldots, j-1\}, \\ 0\,, & \mathcal{I}_{j-1} \text{ such that } C(\boldsymbol{p}_l) \geq \alpha_{l,0} \text{ for } l \in \{1, 2, \ldots, j-1\}, \\ \sum\limits_{l=j}^{J} Pr\left(C(\boldsymbol{p}_{l'}) \in (\alpha_{l,1'}, \alpha_{l,0'}), l' < l;\, C(\boldsymbol{p}_l) \leq \alpha_{l,1} \mid \mathcal{I}_{j-1}\right), & \text{otherwise.} \end{cases} \tag{6.27}$$

The next result presents the weighting vectors that we suggest to use for the stage-wise linear combination z and t tests.

Theorem 6.4.1. *Under* (6.1) *and* (6.22)*, the $j-$th stage predictive power, b_{z_j}, $j = 1, 2, \ldots, J$, of the $J-$stage z test in* (6.27) *is maximized with respect to the weighting vector $\boldsymbol{w}_j$ if and only if $\boldsymbol{w}_j$ is proportional to*

$$\boldsymbol{w}_{z_j^+} = \boldsymbol{\Sigma}^{-1}\boldsymbol{m}_{j-1}. \tag{6.28}$$

Under (6.1) *and* (6.22) *and for* $n_{(j-1)} \to \infty$*, the* $j-$*th stage predictive power,* b_{t_j}*,* $j = 1, 2, \ldots, J$*, of the* $J-$*stage* t *test in* (6.27) *is maximized with respect to the weighting vector* $\boldsymbol{w}_j$ *if and only if* $\boldsymbol{w}_j$ *is proportional to*

$$\boldsymbol{w}_{t_j^+} = \boldsymbol{S}_{j-1}^{-1} \boldsymbol{m}_{j-1}, \tag{6.29}$$

where $\boldsymbol{m}_j$*,* $\boldsymbol{S}_j$ *as in* (6.24) *and* (6.25)*, respectively.*

Proof. The z statistic of the $j-$th stage, $j = 1, 2, \ldots, J$, can be written as

$$z_j = \bar{\theta}_j + e, \qquad e \sim N(0, 1).$$

Under (6.1), (6.22), $(\theta_j \mid \mathcal{I}_{j-1})$ is normally distributed with mean

$$\hat{\theta}_j = \boldsymbol{w}_j^T \boldsymbol{m}_{j-1} / \sigma_j$$

and variance $(n_0 + n_{(j-1)})^{-1}$. Thus, $(z_j \mid \mathcal{I}_{j-1}) \sim N\left(\hat{\theta}_j \sqrt{n}_j, 1 + n_j / \left(n_0 + n_{(j-1)}\right)\right)$. The result is then proved using theorem 6.3.1 and following the same steps as in 4.3.1 where θ_j is replaced by $\hat{\theta}_j$.

For the $t-$test, we wish to compute the asymptotic, $n_{(j-1)} \to \infty$, distribution of $t_j \mid \mathcal{I}_{j-1}$. By Bayes' rule, under (6.1) and (6.22), we have that,

$$(\boldsymbol{x}_{ij} \mid \mathcal{I}_{j-1}) \sim t_K \left(\nu_0 + n_{(j-1)} - K + 1, \boldsymbol{m}_{j-1}, c\boldsymbol{S}_{j-1}\right), \;\; i = 1, 2, \ldots, n_j$$

where $c = (n_0 + n_{(j-1)} + 1)\left((n_0 + n_{(j-1)})(\nu_0 + n_{(j-1)} - K + 1)\right)^{-1}$. Hence for $n_{(j-1)} \to \infty$,

$$(\boldsymbol{x}_{ij} \mid \mathcal{I}_{j-1}) \sim N_K \left(\boldsymbol{m}_{j-1}, \boldsymbol{S}_{j-1}\right).$$

Using the last result and corollary 7.2.3 in Anderson [1984] we have that

$$(\bar{y}_j | \mathcal{I}_{j-1}) \sim N\left(\boldsymbol{w}_j^T m_{j-1}, \boldsymbol{w}_j^T \boldsymbol{S}_{j-1} \boldsymbol{w}_j / n_j\right). \tag{6.30}$$

From the same result we have that

$$\left(\boldsymbol{S}_{\boldsymbol{x}_j}|\mathcal{I}_{j-1}\right) \sim W_{K\times K}\left(n_j-1, \boldsymbol{S}_{j-1}/(n_j-1)\right),$$

and hence by proposition 3.4.2 in Mardia et al. [1979]

$$\left(\frac{s_j^2}{\boldsymbol{w}_j^T\boldsymbol{S}_{j-1}\boldsymbol{w}_j/(n_j-1)}|\mathcal{I}_{j-1}\right) \sim \chi^2_{n_j-1}, \text{ independent of } \bar{y}_j. \tag{6.31}$$

From (6.30) and (6.31), it follows that the t statistic in (6.4) can be written as

$$t_j \mid \mathcal{I}_{j-1} = \frac{Z+\hat{\vartheta}_{j-1}\sqrt{n_j}}{\sqrt{X^2/(n_j-1)}},$$

where $Z \sim N(0,1)$ and $X^2 \sim \chi^2_{n_j-1}$. That is, $t_j \mid \mathcal{I}_{j-1}$ is approximately non-central t distributed with non-centrality parameter $\hat{\vartheta}_{j-1}\sqrt{n_{j-1}}$, where

$$\hat{\vartheta}_{j-1} = \frac{\boldsymbol{w}_{j-1}^T\boldsymbol{m}_{j-1}}{\sqrt{\boldsymbol{w}_{j-1}^T\boldsymbol{S}_{j-1}\boldsymbol{w}_{j-1}}}$$

and n_j-1 degrees of freedom, as $n_{(j-1)} \to \infty$. By replacing θ_j with $\hat{\vartheta}_j$ in the proof of proposition 6.4.1, it follows that, for $n_{(j-1)} \to \infty$, the predictive power function b_{t_j} in (6.27) is maximized with respect to the weighting vector, $\boldsymbol{w}_j$, if $\boldsymbol{w}_j$ is proportional to $\boldsymbol{w}_{t_j^+}$ in (6.29). □

We refer to the $J-$stage linear combination z and t tests with weighting vectors $\boldsymbol{w}_{z_j^+}$ and $\boldsymbol{w}_{t_j^+}$ as the adaptive z_{AD}^+ and t_{AD}^+ tests, respectively.

We can easily prove that these tests satisfy the conditional invariance principle and they control the type I error. We next prove type I error control for the two-stage adaptive z_{AD}^+ and t_{AD}^+ tests, while results for $J-$stage tests follow completely analogously. The main argument is that the weighting vectors $\boldsymbol{w}_{z_j^+} \mid \mathcal{I}_{j-1}$ and $\boldsymbol{w}_{t_j^+} \mid \mathcal{I}_{j-1}$ are fixed and thus, under H_0, the conditional distributions of the stage-wise z_j^+ and t_j^+ statistics are standard normal $N(0,1)$ and Student's t_{n_j-1}

distributed and the corresponding $p-$values uniformly distributed, $U(0,1)$.

Type I error control

For the two-stage z^+_{AD}, t^+_{AD} tests, it is sufficient to show that, under H_0, the p values of the first and second stage, respectively, are

$$p_1, p_2 \mid \mathcal{I}_0 \sim U(0,1), \qquad \text{independent},$$

which implies that, if the critical values $\alpha_{1,0}, \alpha_{1,1}, \alpha_{2,1}$ satisfy the type I error rate equation in (6.10), the type I error rate is controlled. For the rest of this section, all the distributions are computer under H_0.

First see that, if the weighting vectors of the z or t statistics in (6.4) are fixed, their p values are uniformly distributed $U(0,1)$ [George and Mudholkar, 1990]. Conditional on, respectively, $\mathcal{I}_0$ and $\mathcal{I}_1 = \{\mathcal{I}_0, \boldsymbol{X}_1\}$ the weighting vectors of the first and second stage of the adaptive z^+_{AD} and t^+_{AD} tests are fixed and thus,

$$(p_1 \mid \mathcal{I}_0) \sim U(0,1), \qquad (p_2 \mid \mathcal{I}_1) \sim U(0,1). \tag{6.32}$$

It is then sufficient to show that $(p_2 \mid \mathcal{I}_0) \sim U(0,1)$ since this implies also that $(p_2 \mid \mathcal{I}_0)$ and $\boldsymbol{X}_1$ are stochastically independent and thus $p_1, p_2 \mid \mathcal{I}_0$ are also independent. Let $g(\cdot)$, $f(\cdot)$, and $\tilde{g}(\cdot; \boldsymbol{X}_1)$ be the density functions of $(p_2 \mid \mathcal{I}_0)$, $\boldsymbol{X}_1$ and $(p_2 \mid \mathcal{I}_1)$, respectively. Note that, by (6.32), $\tilde{g}(p_2; \boldsymbol{X}_1) = 1$, $p_2 \in [0,1]$ and thus the result follows from

$$g(p_2) = \int \tilde{g}(p_2; \boldsymbol{X}_1) f(\boldsymbol{X}_1) \mathrm{d}\boldsymbol{X}_1 = \int f(\boldsymbol{X}_1) \mathrm{d}\boldsymbol{X}_1 = 1, \quad p_2 \in [0,1],$$

which implies that $(p_2 \mid \mathcal{I}_0) \sim U(0,1)$.

Connections to other linear combination tests

The z^+ and t^+ tests developed in chapter 4 are special cases of the adaptive z^+_{AD} and t^+_{AD} tests. For this, the pilot study is considered as the first stage of the study. The z^+ and t^+ tests are thus two-stage tests with $(\alpha_{1,1}, \alpha_{1,0}) = (0, 1)$, that is, no early stopping is permitted, and $C(\boldsymbol{p}_2) = p_2$. The weighting vector used in z^+ and t^+ tests are, under this representation, equal to $\boldsymbol{w}_{z_2^+}$ and $\boldsymbol{w}_{t_2^+}$, respectively, constructed based on prior information and first stage (pilot) data.

Furthermore, linear combination z and t tests with fixed weighting vectors, such as O'Brien's OLS and GLS (for Σ unknown) tests [O'Brien, 1984], can be implemented under the adaptive design by setting $n_0 \gg n_T$ which effectively sets the weighting vector equal to the first stage weighting vector $\boldsymbol{w}_{z_1^+}$. Alternatively, for group sequential linear combination z and t tests with fixed weighting vectors, one may consider the methodology described in section 5.3.1 (see pp. 72-73).

Two-sided and one-sided p-values

In the proposed z^+_{AD}, t^+_{AD} as well as z^+, t^+ tests, the weighting vectors are allowed to be in any direction the prior information and observed data suggests. Therefore, they are not necessarily restricted to be positive although this can be attained by manipulation. This approach is motivated by neuroimaging in which, as we explained earlier, contrasting effects are often exhibited and are of interest to investigators.

Due to this approach, it is natural to consider two-sided p-values rejecting the null hypothesis H_0 for large absolute values of the z or t statistics regardless of their sign. This suggests that only the direction of the weighting vector or, equivalently, the effect structure is of interest and not the sign.

However, since in our methods the sign of the weighting vector in addition to the direction is chosen, one may consider one-sided p-values in order to improve power. Specifically, the z^+_{AD} and t^+_{AD} tests described earlier can be implemented using the p-values $p_{z_j} = \Phi(-z_j)$ or $p_{t_j} = \Psi(t_j)$, respectively. Power improvement

will be attained if the sign of the weighting vector is correctly chosen.

In adaptive testing methodology, it is often preferable to use one-sided p-values since for univariate stage-wise tests this prevents rejections based on effects with contrasting signs. The issue is more complicated in multivariate testing and if one wish to prevent such situations should consider one-sided tests (see section 3.3.2). Primarily due to our motivating application, in this thesis, we mainly consider tests with two-sided p-values.

6.5 Conclusions

The methodology developed in this chapter provides an important generalisation of the testing procedures described in chapter 4. The issues arising with external pilot studies are overcome with a more efficient use of the pilot data. The latter are here used not only for selecting the weighting vector, but also for testing. This potentially leads to a considerable improvement to the power performance of the test, particularly in the case where the first stage weighting vector is close to optimal. In addition, the tests are further generalised to allow for more than two stages, with the weighting vector being adapted to observed data at every interim analysis. Moreover, early stopping of the study, which is particularly important in medical settings, is permitted in these designs.

The tests control type I error under general conditions and they are optimal with respect to predictive power. They can be used in situations where a pre-specification of the adaptation rules is required. In the next chapter, we derive a power characterisation of linear combination tests, which is then used in chapter 8 to perform an extensive power analysis of these tests, including comparisons to alternative global tests. The test statistic adaptation is expected to improve power performance in cases where the initial statistic is far from optimal, while it may lead to lose of power if the latter is close to optimal.

Chapter 7

Power characterisation for linear combination tests

7.1 Introduction

In order to study the performance of a test, we primarily need to explore the relation between its power function and the parameters defining the statistical model and the design of the study. Design parameters include the sample size(s), the critical value(s) of the test and the sample allocation to different observation groups. In linear combination tests, another key design parameter is the weighting vector of the linear combinations. This defines the region we decide to search for effects, or, in other words, the "targeted" region of the test.

The critical values and the sample size(s) of the test are scalar and therefore it is straightforward to visualize power even across all their possible values (for example using simulations). Their relation to power can then be relatively easily described and understood. In univariate settings, this is also the case for the model parameters. For example, to study the relation between power and treatment effect, we can plot power versus the standardized effect $\theta = \mu/\sigma$ for a fixed sample size.

However, in the multivariate setting, the model parameters and other multi-

variate design parameters can be high dimensional and therefore it is not practically feasible to visualize power over all the possible values of these parameters. This makes our understanding to power behavior very limited. For Hotelling's T^2 test, this problem is alleviated since the relation of power to the model parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be summarised by the Mahalanobis distance Δ. The Mahalanobis distance is scalar and measures the strength of the global treatment effect. Thus, in T^2, the relation of power to the size of the global effect can be easily studied.

However, such summaries, characterizing the relation of power to model and design parameters, are not yet identified for the one-sided tests [Follmann, 1996; Tang et al., 1989b] nor for the O'Brien-type and Läuter's linear combination tests [Läuter, 1996; O'Brien, 1984]. Power analysis for these tests is then typically restricted to simulations of the power for a limited range of mean and covariance structures.

As we discuss in earlier chapters, in addition to the strength of the global effect, an important factor for the power of every multivariate test, including T^2, is the direction of the effect. For linear combination or other "directional" tests, in particular, power is substantially affected by the distance between the effect direction and the targeted direction. Despite that various authors acknowledge this fact (see for example Follmann [1996]; Logan and Tamhane [2004]; Pocock et al. [1987]; Tang et al. [1993, 1989b]), a measure of the latter distance is yet to be established.

In the following, we encounter this problem in the context of linear combination tests and we provide a solution. We first consider the case of $J-$stage linear combination z and t tests with fixed weighting vectors which, apart from providing a method for performing simple power analysis of tests such as O'Brien's OLS test (for earlier work see Logan and Tamhane [2004]; Pocock et al. [1987]; Tang et al. [1993]), also provides the intuition for the results considering the z^+_{AD} and t^+_{AD} tests.

For the remainder of this chapter, the critical values and sample sizes of the $J-$stage tests are assumed to be fixed and described by the design vector $\boldsymbol{d} =$

$(\alpha_{0,1}, \alpha_{0,2}, \ldots, \alpha_{0,J-1}, \alpha_{1,1}, \alpha_{1,2}, \ldots, \alpha_{1,J}, \nu_0, n_0, n_1, \ldots, n_J)$. Our primary interest is in the model parameters and the weighting vector.

7.2 $J-$stage z and t tests

To compute the power of the $J-$stage z and t tests with fixed weighting vectors $\boldsymbol{w}_j = \boldsymbol{w}$, it is sufficient to know the design vector $\boldsymbol{d}$, as well as the stage-wise location parameters θ_j in (6.6) (page 87). The latter parameters, for fixed $\boldsymbol{w}_j = \boldsymbol{w}$, $j = 1, 2, \ldots, J$, are also fixed and can be re-expressed as

$$\theta = \frac{\boldsymbol{w}^T\boldsymbol{\mu}}{\sqrt{\boldsymbol{w}^T\boldsymbol{\Sigma}\boldsymbol{w}}} = \frac{\tilde{\boldsymbol{w}}^T\tilde{\boldsymbol{\omega}}^*}{\|\tilde{\boldsymbol{w}}\|} = \| \tilde{\boldsymbol{\omega}}^* \| \cos\phi_{\boldsymbol{w}}, \quad \phi_{\boldsymbol{w}} = ang(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{\omega}}^*), \tag{7.1}$$

where $\phi_{\boldsymbol{w}}$ denotes the angle, in measured degrees at the origin, between respectively the standardized selected and optimal weighting vectors

$$\tilde{\boldsymbol{w}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{w}, \quad \tilde{\boldsymbol{\omega}}^* = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\omega}^* = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{\mu}. \tag{7.2}$$

The standardized optimal weighting vector, $\tilde{\boldsymbol{\omega}}^*$, expresses the standardized, to the covariance matrix, multivariate treatment effect. This standardization eliminates the variances and correlations of the local outcomes. The vector $\tilde{\boldsymbol{\omega}}^*$ is a multivariate generalization of the univariate ($K = 1$) standardized treatment effect μ/σ.

Considering the weighting vector selection problem, the first equation in (7.1) implies that a weighting vector that increases the mean and/or decreases the variance of the linear combination gives higher power. The ambiguity in the latter expression becomes clearer by the standardization in the second equation which implies that the weighting vector selection can be expressed as a process of learning the direction of the standardized optimal weighting vector $\tilde{\boldsymbol{\omega}}^*$, that is, the standardized treatment effect direction.

The last equation in (7.1) establishes two scalar measures which are sufficient to determine power. The first is the magnitude of $\tilde{\boldsymbol{\omega}}^*$,

$$\|\tilde{\boldsymbol{\omega}}^*\| = \left(\boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}\right)^{1/2} = \Delta, \tag{7.3}$$

which is equal to the Mahalanobis distance. As we discussed earlier, Mahalanobis distance is a global measure of the strength of the treatment effect. The second, $\cos\phi_{\boldsymbol{w}}$, with $\phi_{\boldsymbol{w}} = ang(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{\omega}}^*)$, is a measure of angular distance between the selected weighting vector and the optimal weighting vector. In other words, the angular distance between the targeted direction and the effect direction.

To summarise, for fixed weighting vectors, the location parameter, θ, is equal to a measure, Δ, of the strength of the treatment effect scaled down by a measure, $\cos\phi_{\boldsymbol{w}}$, of the distance between the effect direction and the targeted direction expressed by the selected weighting vector. Importantly, as the next theorem states, the latter measures determine the relation of the model parameters and the weighting vector to power.

Theorem 7.2.1. *The design vector $\boldsymbol{d}$, the Mahalanobis distance Δ in* (7.3) *and the angle $\phi_{\boldsymbol{w}}$ between the vectors $\tilde{\boldsymbol{\omega}}^*$ and $\tilde{\boldsymbol{w}}$ in* (7.2) *are sufficient to determine the power function β of the $J-$stage linear combination z and t tests with fixed weighting vectors $\boldsymbol{w}_j = \boldsymbol{w}$.*

Proof. By (6.4), (6.6) and (6.9) (pp. 86, 87, 88), the values of the design vector $\boldsymbol{d}$ and the parameter θ_j, $j = 1, 2, \ldots, J$ are sufficient to compute the power function of the $J-$stage z and t tests with $\boldsymbol{w}_j = \boldsymbol{w}$, $j = 1, 2, \ldots, J$. The result then follows from the last equation in (7.1) and equation (7.3). □

It is worth noting that the above results stand for the $J-$stage linear combination z and t tests (fixed weighing vector), as described in the previous chapter, but they also apply to the group sequential linear combination z_{GS} and t_{GS} tests,

as described in section 5.3.1 (page 72). The description derived in the last theorem is sufficient to compute power and, importantly, is based on easily interpretable summaries which are scalar (see table 7.1).

Table 7.1: Model parameters, weighting vector and their dimension for the linear combination z, t, z_{GS} and t_{GS} tests.

parameters	dimension
$\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{w}$	$\left(K^2 + 5K\right)/2$
$\tilde{\boldsymbol{\omega}}^*, \tilde{\boldsymbol{w}}$	$2K$
$\Delta, \phi_{\boldsymbol{w}}$	2

This allows us to perform power analysis of the single-stage z, t and group sequential z_{GS} and t_{GS} tests in a simple way potentially covering the whole design space.

7.3 $J-$stage z^+_{AD} test

The sequential adaptation of the weighting vector increases the complexity within the relation between power and the design and model parameters. However, following similar methodology as above, analogous results can be derived.

For this we use two steps, the first of which involves standardizing the procedure similarly to (7.1), and the second establishing a rotation invariance property of the power function. The next lemma is a direct consequence of the standardization step summarizing $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{m}_0$ to the vectors $\tilde{\boldsymbol{\omega}}^*$ and

$$\tilde{\boldsymbol{w}}_{z_1^+} \;=\; \boldsymbol{\Sigma}^{1/2}\boldsymbol{w}_{z_1^+} \;=\; \boldsymbol{\Sigma}^{-1/2}\boldsymbol{m}_0. \tag{7.4}$$

Lemma 7.3.1. *The design vector $\boldsymbol{d}$, the standardized optimal weighting vector $\tilde{\boldsymbol{\omega}}^*$ in* (7.2) *and the standardized first-stage weighting vector $\tilde{\boldsymbol{w}}_{z_1^+}$ in* (7.4) *are sufficient to determine the power function $\beta_{z^+_{AD}}$.*

Proof. By (6.4) and (6.9) (pp. 86, 88), the values of the design vector $\boldsymbol{d}$ and the parameter $\theta_{z_j^+}$ in (7.5), $j = 1, 2, \ldots, J$, are sufficient to compute the power function

of the z_{AD}^{+} tests. Furthermore, we can write the location parameter, $\theta_{z_j^+}$, and the standardized weighting vector, $\tilde{\boldsymbol{w}}_{z_j^+} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{m}_{j-1}$, $j = 1, 2, \ldots, J$, of the z_{AD}^{+} test as

$$\theta_{z_j^+} = \frac{\tilde{\boldsymbol{w}}_{z_j^+}^T \tilde{\boldsymbol{\omega}}^*}{\|\tilde{\boldsymbol{w}}_{z_j^+}\|}, \quad \tilde{\boldsymbol{w}}_{z_j^+} = \frac{n_0 \tilde{\boldsymbol{w}}_{z_1^+} + n_{(j-1)} \tilde{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j-1)}}}{n_0 + n_{(j-1)}}, \tag{7.5}$$

where

$$\tilde{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j)}} = \boldsymbol{\Sigma}^{-1/2}\bar{\boldsymbol{x}}_{(j)} \sim N_K(\tilde{\boldsymbol{\omega}}^*, \boldsymbol{I}/n_{(j)}). \tag{7.6}$$

By equations (7.5) and (7.6), we have that $\tilde{\boldsymbol{\omega}}^*$ in (7.2) and $\tilde{\boldsymbol{w}}_{z_1^+}$ in (7.4) are sufficient to determine $\theta_{z_j^*}$, $j = 1, 2, \ldots, J$, and the result follows. □

The last lemma and particularly equation (7.5) imply that the adaptive selection of the weighting vectors can be re-expressed as a procedure of adaptive estimation of the direction of $\tilde{\boldsymbol{\omega}}^*$. Furthermore, under this standardization, we can proceed to the rotation-invariance step which results in the next lemma.

Lemma 7.3.2. *The power function $\beta_{z_{AD}^+}$ is invariant to rotations of the weighting vector $\tilde{\boldsymbol{w}}_{z_1^+}$ in* (7.5) *around the optimal weighting vector $\tilde{\boldsymbol{\omega}}^*$ in* (7.2).

Proof. Let $\boldsymbol{R}$ be a rotation matrix with rotation axis $\tilde{\boldsymbol{\omega}}^*$ ($\boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{I}$, $\boldsymbol{R}\tilde{\boldsymbol{\omega}}^* = \tilde{\boldsymbol{\omega}}^*$). We define $\dot{\boldsymbol{w}}_{z_1^+} = \boldsymbol{R}\tilde{\boldsymbol{w}}_{z_1^+}$ the rotated first stage weighting vector (note that $\|\dot{\boldsymbol{w}}_{z_1^+}\| = \|\tilde{\boldsymbol{w}}_{z_1^+}\|$, $ang(\tilde{\boldsymbol{\omega}}^*, \dot{\boldsymbol{w}}_{z_1^+}) = ang(\tilde{\boldsymbol{\omega}}^*, \tilde{\boldsymbol{w}}_{z_1^+})$). We can write $\tilde{\boldsymbol{w}}_{z_j^+}$ in (7.5), $j = 1, 2, \ldots, J$, as

$$\tilde{\boldsymbol{w}}_{z_j^+} = \boldsymbol{R}^T\dot{\boldsymbol{w}}_{z_j^+}, \quad \dot{\boldsymbol{w}}_{z_j^+} = \frac{n_0 \dot{\boldsymbol{w}}_{z_1^+} + n_{(j-1)} \dot{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j-1)}}}{n_0 + n_{(j-1)}}, \qquad (n_{(0)} = 0)$$

where $\dot{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j)}} = \boldsymbol{R}\tilde{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j)}}$ and $\tilde{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j)}}$ as in (7.6). Therefore, the stage-wise test statistics z_j^+, which along with the critical values fully describe the rejection region of the z_{AD}^{+} test, can be re-written as

$$z_j^+ = \frac{\boldsymbol{w}_{z_j^+}^T \bar{\boldsymbol{x}}_j}{\sqrt{\boldsymbol{w}_{z_j^+}^T \boldsymbol{\Sigma} \boldsymbol{w}_{z_j^+}/n_j}} = \frac{\tilde{\boldsymbol{w}}_{z_j^+}^T \tilde{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_j}}{\|\tilde{\boldsymbol{w}}_{z_j^+}\|/\sqrt{n_j}} = \frac{\dot{\boldsymbol{w}}_{z_j^+}^T \boldsymbol{R}\tilde{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_j}}{\|\dot{\boldsymbol{w}}_{z_j^+}\|/\sqrt{n_j}} = \frac{\dot{\boldsymbol{w}}_{z_j^+}^T \dot{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_j}}{\|\dot{\boldsymbol{w}}_{z_j^+}\|/\sqrt{n_j}},$$

where $\dot{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_j} = \boldsymbol{R}\tilde{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_j}$ and $\tilde{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_j} = \boldsymbol{\Sigma}^{-1/2}\bar{\boldsymbol{x}}_j$. Therefore, the rejection region of the $J-$stage z_{AD}^{+} test after the rotation is equal to the initial but with the first stage standardized weighting vector $\tilde{\boldsymbol{w}}_{z_1^+}$ replaced by $\dot{\boldsymbol{w}}_{z_1^+}$ and the standardized observation vectors $\tilde{\boldsymbol{w}}_{\boldsymbol{x}_{ij}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_{ij}$ replaced by $\dot{\boldsymbol{w}}_{\boldsymbol{x}_{ij}} = \boldsymbol{R}\tilde{\boldsymbol{w}}_{\boldsymbol{x}_{ij}}$, $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, J$. Importantly, $\dot{\boldsymbol{w}}_{\boldsymbol{x}_{ij}}$ has the same distribution, $N_K(\tilde{\boldsymbol{\omega}}^*, \boldsymbol{I})$, as $\tilde{\boldsymbol{w}}_{\boldsymbol{x}_{ij}}$ and thus the result follows. □

The result of lemma 7.3.2 can be described, in other words, as that the change to the rejection region of the z_{AD}^{+} test produced by rotating $\tilde{\boldsymbol{w}}_{z_1^+}$ around $\tilde{\boldsymbol{\omega}}^*$, that is, replacing $\tilde{\boldsymbol{w}}_{z_1^+}$ by $\dot{\boldsymbol{w}}_{z_1^+} = \boldsymbol{R}\tilde{\boldsymbol{w}}_{z_1^+}$ (where $\boldsymbol{R}$ a rotation matrix with axis $\tilde{\boldsymbol{\omega}}^\star$), does not affect power. To see this, first note that the new rejection region, after the rotation of $\tilde{\boldsymbol{w}}_{z_1^+}$, is simply a rotation of the initial rejection region. That is, for each point say $\tilde{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j)}}$ in the initial rejection region, we can find a unique point, say $\dot{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j)}}$, in the rotated rejection region such that $\dot{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j)}} = \boldsymbol{R}\tilde{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j)}}$. The power is unchanged by this rotation, because the symmetrical Gaussian distribution of the observations $\tilde{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j)}} \sim N_K(\tilde{\boldsymbol{\omega}}^*, \boldsymbol{I}/n_{(j)})$ remains unchanged under the rotation. This implies the result which is that the likelihood of the rejection region, i.e. the power of the z_{AD}^{+} test, remains the same after the rotation.

The next theorem is direct consequence of lemmas 7.3.1 and 7.3.2.

Theorem 7.3.1. *The design vector* $\boldsymbol{d}$*, the Mahalanobis distance* Δ *in* (7.3) *and the angle* $\phi_{z_1^+} = ang(\tilde{\boldsymbol{\omega}}^*, \tilde{\boldsymbol{w}}_{z_1^+})$ *between the vectors* $\tilde{\boldsymbol{\omega}}^*$ *in* (7.2) *and* $\tilde{\boldsymbol{w}}_{z_1^+}$ *in* (7.4) *are sufficient to determine the power function* $\beta_{z_{AD}^+}$.

Proof. By lemma 7.3.1, the design vector $\boldsymbol{d}$ and the vectors $\tilde{\boldsymbol{\omega}}^*$ and $\tilde{\boldsymbol{w}}_{z_1^+}$ are sufficient to determine the power function $\beta_{z_{AD}^+}$. Further, lemma 7.3.2, by rotation properties, implies that for any value of $\tilde{\boldsymbol{w}}_{z_1^+}$ such that $\phi_{z_1^+}$ remains fixed, the power $\beta_{z_{AD}^+}$ test is invariant. The result then follows immediately. □

The above theorem states that the dependence of the power function on the model parameters and their prior estimates is described by simply a measure of

the strength of the treatment effect and a measure of distance between the model parameters, determinging the effect direction, and their prior estimates, determining the targeted direction. These easily interpretable measures, which provide a sufficient description of power, are also scalar (see table 7.2). Therefore this power characterisation, apart from providing great inside on the behavior of z^+ and z^+_{AD} tests, allows us to perform simple power analysis potentially covering the whole design space.

Table 7.2: Model and prior parameters of the z^+ and z^+_{AD} tests and their dimension.

parameters	dimension
$\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{m}_0$	$\left(K^2+5K\right)/2$
$\tilde{\boldsymbol{\omega}}^*, \tilde{\boldsymbol{w}}_{z_1^+}$	$2K$
$\Delta, \phi_{z_1^+}$	2

7.4 $J-$stage t^+_{AD} test

The need to estimate the unknown $\boldsymbol{\Sigma}$ increases substantially the dimension and the complexity of the design space. In particular, the sequential estimation of $\boldsymbol{\Sigma}$, in addition to $\boldsymbol{\mu}$, to obtain the weighting vectors $\boldsymbol{w}_{t_j^+} = \boldsymbol{S}_{j-1}^{-1}\boldsymbol{m}_{j-1}$ in (6.29) (page 99), implies that the power analysis needs to account for both estimation procedures. For this, we write the weighting vector $\tilde{\boldsymbol{w}}_{t_j^+} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{w}_{t_j^+}$, $j = 1, 2, \ldots, J$ as

$$\tilde{\boldsymbol{w}}_{t_j^+} = \boldsymbol{D}_j^{-1}\tilde{\boldsymbol{w}}_{z_j^+}, \qquad \boldsymbol{D}_j = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}_{j-1}\boldsymbol{\Sigma}^{-1/2} \tag{7.7}$$

and $\tilde{\boldsymbol{w}}_{z_j^+}$ the $j-$th standardized weighting vector of the z^+_{AD} test in (7.5). Here the $\boldsymbol{\Sigma}-$deviation matrix $\boldsymbol{D}_j$ is a measure of deviation of the estimate $\boldsymbol{S}_{j-1}$ in (6.25) (see page 98) from the parameter $\boldsymbol{\Sigma}$. The weighting vector $\tilde{\boldsymbol{w}}_{t_j^+}$ is then written as a product of the inverse of the matrix $\boldsymbol{D}_j$, that accounts for the estimation of $\boldsymbol{\Sigma}$, and the vector $\tilde{\boldsymbol{w}}_{z_j^+}$ which accounts for the estimation of $\boldsymbol{\mu}$, the latter taking $\boldsymbol{\Sigma}$ as known.

We next follow the same steps as in section 7.3 for deriving the power characterization of the t^+_{AD} test. The standardization step results in the next lemma summarizing $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and their prior estimates $\boldsymbol{m}_0$ and $\boldsymbol{S}_0$ to the vectors $\tilde{\boldsymbol{\omega}}^*$, $\tilde{\boldsymbol{w}}_{z_1^+}$ and the matrix $\boldsymbol{D}_1$ that have clear interpretation.

Lemma 7.4.1. *The design vector* $\boldsymbol{d}$*, the matrix* $\boldsymbol{D}_1$ *in* (7.7) *and the vectors* $\tilde{\boldsymbol{\omega}}^*$ *in* (4.11) *and* $\tilde{\boldsymbol{w}}_{z_1^+}$ *in* (7.5) *are sufficient to determine the power function* $\beta_{t^+_{AD}}$*.*

Proof. By (6.4) and (6.9), the design vector $\boldsymbol{d}$ and $\theta_{t_j^+}$ in (7.9), $j = 1, 2, \ldots, J$, are sufficient to compute the power function of the t^+_{AD} tests. The location parameter, $\theta_{t_j^+}$, of the t_j^+ statistic can be written as

$$\theta_{t_j^+} = \frac{\tilde{\boldsymbol{w}}^T_{t_j^+}\tilde{\boldsymbol{\omega}}^*}{\|\tilde{\boldsymbol{w}}_{t_j^+}\|} = \frac{\tilde{\boldsymbol{w}}^T_{z_j^+}\boldsymbol{D}_j^{-1}\tilde{\boldsymbol{\omega}}^*}{\|\boldsymbol{D}_j^{-1}\tilde{\boldsymbol{w}}_{z_j^+}\|}, \tag{7.8}$$

where $\tilde{\boldsymbol{w}}^T_{z_j^+}$ can be written as the weighted average in (7.5) and the $\boldsymbol{\Sigma}-$deviation matrix $\boldsymbol{D}_j$ as

$$\boldsymbol{D}_j = \boldsymbol{D}_1 + \nu_{(j-1)}\boldsymbol{S}_{\tilde{\boldsymbol{w}}_{\boldsymbol{x}_{(j-1)}}} + \frac{n_0 n_{(j-1)}}{n_0 + n_{(j-1)}}\left(\tilde{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j-1)}} - \tilde{\boldsymbol{w}}_{z_1^+}\right)\left(\tilde{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j-1)}} - \tilde{\boldsymbol{w}}_{z_1^+}\right)^T, \tag{7.9}$$

where $\boldsymbol{S}_{\tilde{\boldsymbol{w}}_{\boldsymbol{x}_{(j)}}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{S}_{\boldsymbol{x}_{(j)}}\boldsymbol{\Sigma}^{-1/2}$ is the covariance matrix of the sample $\tilde{\boldsymbol{w}}_{\boldsymbol{x}_{il}}$, $i = 1, 2, \ldots, n_l$, $l = 1, 2, \ldots, j$. Here, importantly, $\tilde{\boldsymbol{w}}_{\boldsymbol{x}_{il}} = \boldsymbol{\Sigma}^{-1/2}\boldsymbol{x}_{il} \sim N_K(\tilde{\boldsymbol{\omega}}^*, \boldsymbol{I})$. Thus, we have that $\tilde{\boldsymbol{\omega}}^*$ in (7.2), $\tilde{\boldsymbol{w}}_{z_1^+}$ in (7.4) and $\boldsymbol{D}_1$ in (7.7) are sufficient to determine $\theta_{t_j^*}$, $j = 1, 2, \ldots, J$, from which the result follows. □

In a similar fashion to the previous section, we next establish the invariance of the power function under certain rotations of the prior estimates. For this, we define $\boldsymbol{V} = [\boldsymbol{v}_1\ \boldsymbol{v}_2\ \ldots\ \boldsymbol{v}_K]$ to be the matrix with the columns being the orthonormal eigenvectors $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_K$ of $\boldsymbol{D}_1$ and $\boldsymbol{\Lambda}_1 = diag(\boldsymbol{\lambda}_1)$ the diagonal matrix with diagonal $\boldsymbol{\lambda}_1 = (\lambda_{11}, \lambda_{12}, \ldots, \lambda_{1K})^T$ the vector of the corresponding eigenvalues $(\lambda_{11} \geq \lambda_{21} \geq \cdots \geq \lambda_{1K} > 0)$. We can then write $\boldsymbol{D}_1 = \boldsymbol{V}\boldsymbol{\Lambda}_1\boldsymbol{V}^T$, $\tilde{\boldsymbol{w}}_{z_1^+} = \boldsymbol{V}\boldsymbol{c}_{z_1^+}$ and

$\tilde{\boldsymbol{\omega}}^* = \boldsymbol{V}\boldsymbol{c}^*$ where

$$c_{z_j^+,k} = \cos(ang(\boldsymbol{v}_k, \tilde{\boldsymbol{w}}_{z_j^+})), \quad c_k^* = \cos\left(ang\left(\boldsymbol{v}_k, \tilde{\boldsymbol{\omega}}^*\right)\right),\ k = 1, 2, \ldots, K. \tag{7.10}$$

The rotation invariance property of the t_{AD}^+ test is described in the next lemma.

Lemma 7.4.2. *The power function $\beta_{t_{AD}^+}$ is invariant to simultaneous rotations of the vector $\tilde{\boldsymbol{w}}_{z_1^+}$ in* (7.5) *and the eigenvectors of the matrix $\boldsymbol{D}_1$ in* (7.7) *around the optimal weighting vector $\tilde{\boldsymbol{\omega}}^*$ in* (7.2)*.*

Proof. Let $\boldsymbol{R}$ be a rotation matrix with rotation axis $\tilde{\boldsymbol{\omega}}^*$ and define $\dot{\boldsymbol{w}}_{z_1^+} = \boldsymbol{R}\tilde{\boldsymbol{w}}_{z_1^+}$, $\dot{\boldsymbol{V}} = \boldsymbol{R}\boldsymbol{V} = [\dot{\boldsymbol{v}}_1\ \dot{\boldsymbol{v}}_2\ \ldots\ \dot{\boldsymbol{v}}_K]$ (note that $c_k^* = \cos(ang(\tilde{\boldsymbol{\omega}}^*, \boldsymbol{v}_k)) = \cos(ang(\tilde{\boldsymbol{\omega}}^*, \dot{\boldsymbol{v}}_k))$, $\boldsymbol{c}_{z_1^+,k} = \cos(ang(\tilde{\boldsymbol{w}}_{z_1^+}, \boldsymbol{v}_k)) = \cos(ang(\dot{\boldsymbol{w}}_{z_1^+}, \dot{\boldsymbol{v}}_k))$, $k = 1, 2, \ldots, K$). Then note that we can write $\tilde{\boldsymbol{w}}_{z_j^+}$ in (7.5) as

$$\tilde{\boldsymbol{w}}_{z_j^+} = \boldsymbol{R}^T\dot{\boldsymbol{w}}_{z_j^+}, \quad \dot{\boldsymbol{w}}_{z_j^+} = \frac{n_0\dot{\boldsymbol{w}}_{z_1^+} + n_{(j-1)}\dot{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j-1)}}}{n_0 + n_{(j-1)}}, \qquad (n_{(0)} = 0), \tag{7.11}$$

and $\boldsymbol{D}_j$ in (7.9) as $\boldsymbol{D}_j = \boldsymbol{R}^T\dot{\boldsymbol{D}}_j\boldsymbol{R}$, where

$$\dot{\boldsymbol{D}}_j = \dot{\boldsymbol{D}}_1 + (n_{(j-1)} - 1)\boldsymbol{S}_{\dot{\boldsymbol{w}}_{\boldsymbol{x}_{(j-1)}}} + \frac{n_0 n_{(j-1)}}{n_0 + n_{(j-1)}}\left(\dot{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j-1)}} - \dot{\boldsymbol{w}}_{z_1^+}\right)\left(\dot{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j-1)}} - \dot{\boldsymbol{w}}_{z_1^+}\right)^T,$$

where $\dot{\boldsymbol{D}}_1 = \dot{\boldsymbol{V}}\boldsymbol{\Lambda}_1\dot{\boldsymbol{V}}^T$ and $\dot{\boldsymbol{w}}_{\boldsymbol{x}_{(j)}}$, $\boldsymbol{S}_{\dot{\boldsymbol{w}}_{\boldsymbol{x}_{(j-1)}}}$ the sample mean and sample covariance matrix of the sample $\dot{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{ij}} = \boldsymbol{R}\tilde{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{ij}}$, $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, J$. Therefore, the weighting vector $\tilde{\boldsymbol{w}}_{t_j^+}$ in (7.7) can be written as

$$\tilde{\boldsymbol{w}}_{t_j^+} = \boldsymbol{R}^T\dot{\boldsymbol{w}}_{t_j^+}, \qquad \dot{\boldsymbol{w}}_{t_j^+} = \dot{\boldsymbol{D}}_j^{-1}\dot{\boldsymbol{w}}_{z_j^+}$$

and hence the stage-wise test statistics, t_j^+ (which, along with the critical values,

fully describe the rejection region of the t^+_{AD} test) can be re-written as

$$t_j^+ = \frac{\tilde{\boldsymbol{w}}^T_{t_j^+}\tilde{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_j}\sqrt{n_j}}{\sqrt{\tilde{\boldsymbol{w}}^T_{t_j^+}\boldsymbol{S}_{\tilde{\boldsymbol{w}}_{\boldsymbol{x}_j}}\tilde{\boldsymbol{w}}_{t_j^+}}} = \frac{\dot{\boldsymbol{w}}^T_{t_j^+}\boldsymbol{R}\tilde{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_j}\sqrt{n_j}}{\sqrt{\dot{\boldsymbol{w}}^T_{t_j^+}\boldsymbol{R}\boldsymbol{S}_{\tilde{\boldsymbol{w}}_{\boldsymbol{x}_j}}\boldsymbol{R}^T\dot{\boldsymbol{w}}_{t_j^+}}} = \frac{\dot{\boldsymbol{w}}^T_{t_j^+}\dot{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_j}\sqrt{n_j}}{\sqrt{\dot{\boldsymbol{w}}^T_{t_j^+}\boldsymbol{S}_{\dot{\boldsymbol{w}}_{\boldsymbol{x}_j}}\dot{\boldsymbol{w}}_{t_j^+}}}. \tag{7.12}$$

Therefore, the rejection region of the t^+_{AD} test after the rotation is equal to the initial but with $\boldsymbol{V}$ replaced by $\dot{\boldsymbol{V}}$, $\tilde{\boldsymbol{w}}_{z_1^+}$ replaced by $\dot{\boldsymbol{w}}_{z_1^+}$ and $\tilde{\boldsymbol{w}}_{\boldsymbol{x}_{ij}}$ replaced by $\dot{\boldsymbol{w}}_{\boldsymbol{x}_{ij}}$. Importantly, $\dot{\boldsymbol{w}}_{\boldsymbol{x}_{ij}}$, $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, J$, have the same distribution, $N_K(\tilde{\boldsymbol{\omega}}^*, \boldsymbol{I})$, as $\tilde{\boldsymbol{w}}_{\boldsymbol{x}_{ij}}$ and thus the result follows. □

The above proof, albeit with the extra complexity, is based on the same idea with the proof of lemma 7.3.2. The next theorem is a direct consequence of lemmas 7.4.1 and 7.4.2.

Theorem 7.4.1. *The design vector* $\boldsymbol{d}$*, the vector of eigenvalues* $\boldsymbol{\lambda}_1$ *of* $\boldsymbol{D}_1$ *in* (7.7) *and the vectors* $\boldsymbol{c}_{z_1^+}$ *and* $\boldsymbol{c}^*$ *in* (7.10) *are sufficient to determine the power function* $\beta_{t^+_{AD}}$.

Proof. By lemma 7.4.1, the design vector $\boldsymbol{d}$ and $\tilde{\boldsymbol{\omega}}^*$, $\tilde{\boldsymbol{w}}_{z_1^+}$ and $\boldsymbol{D}_1$ are sufficient to determine the power function $\beta_{t^+_{AD}}$. Furthermore, lemma 7.4.2, by rotation properties, implies that for any value of $\tilde{\boldsymbol{w}}_{z_1^+}$ and $\boldsymbol{D}_1$ such that the vectors $\boldsymbol{c}^*$, $\boldsymbol{c}_{z_1^+}$ in (7.10) and $\boldsymbol{\lambda}_1$ remain fixed, the power of the t^+_{AD} remains fixed. The result then follows. □

As we can see in table 7.3, the last result reduces the dimension of the design space of the t^+_{AD} test substantially. While the design space, due to the covariance matrix estimation, still depends on K, it is reduced from order K^2 to order K. It is useful to note that these results apply to the t^+ as a special case to the t^+_{AD} test.

Table 7.3: Model and prior parameters of the t^+ and t^+_{AD} tests and their dimension

parameters	dimension
$\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{m}_0, \boldsymbol{S}_0$	$K^2 + 3K$
$\tilde{\boldsymbol{\omega}}^*, \tilde{\boldsymbol{w}}_{z_1^+}, \boldsymbol{D}_1$	$\frac{K^2+5K}{2}$
$\boldsymbol{c}^*, \boldsymbol{c}_{z_1^+}, \boldsymbol{\lambda}_1$	$3K$

Furthermore, this reduction provides an understanding of how the selection of the weighting vector affects power. This becomes clearer if we consider that $\theta_{t_j^+}$ in (7.9) can be written as

$$\theta_{t_j^+} = \frac{\boldsymbol{c}_{z_j^+}^T \boldsymbol{\Lambda}_j^{-1} \boldsymbol{c}^*}{\|\boldsymbol{\Lambda}_j^{-1} \boldsymbol{c}_{z_j^+}\|}, \; j = 1, 2, \ldots, J,$$

where

$$\boldsymbol{c}_{z_j^+} = \frac{n_0 \boldsymbol{c}_{z_1^+} + n_{(j-1)} \boldsymbol{c}_{\bar{\boldsymbol{x}}_{(j-1)}}}{n_0 + n_{(j-1)}},$$

and

$$\boldsymbol{\Lambda}_j = \boldsymbol{\Lambda}_1 + \nu_{(j-1)} \boldsymbol{S}_{\boldsymbol{c}_{\boldsymbol{x}_{(j-1)}}} + \frac{n_0 n_{(j-1)}}{n_0 + n_{(j-1)}} (\boldsymbol{c}_{\bar{\boldsymbol{x}}_{(j-1)}} - \boldsymbol{c}_{z_1^+})(\boldsymbol{c}_{\bar{\boldsymbol{x}}_{(j-1)}} - \boldsymbol{c}_{z_1^+})^T.$$

Here, $\boldsymbol{c}_{\bar{\boldsymbol{x}}_{(j)}}$ and $\boldsymbol{S}_{\boldsymbol{c}_{\boldsymbol{x}_{(j)}}}$ are the sample mean and sample covariance matrix of the transformed observation vectors $\boldsymbol{c}_{\boldsymbol{x}_{(j)}} = [\boldsymbol{c}_{\boldsymbol{x}_1} \boldsymbol{c}_{\boldsymbol{x}_2} \ldots \boldsymbol{c}_{\boldsymbol{x}_j}]$ with $\boldsymbol{c}_{\boldsymbol{x}_l}$, $l = 1, 2, \ldots, j$, the matrix with columns $\boldsymbol{c}_{\boldsymbol{x}_{il}} = \boldsymbol{V}_1^T \tilde{\boldsymbol{w}}_{\boldsymbol{x}_{il}} \sim N_K(\boldsymbol{c}^*, \boldsymbol{I})$, $i = 1, 2, \ldots, n_j$. The last expressions show that the distance of the prior estimates $\boldsymbol{m}_0$, $\boldsymbol{S}_0$ to the model parameters $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ can be expressed by the distances of the vectors $\boldsymbol{c}_{z_1^+}$ and $\boldsymbol{\lambda}_1^{-1} = (1/\lambda_{11}, \ldots, 1/\lambda_{1K})^T$ to $\boldsymbol{c}^*$, the latter directly reflected to power through $\theta_{t_j^+}$ (see the next chapter for more information).

In the special case of the first stage $\boldsymbol{\Sigma}-$deviation matrix being proportional to the identity matrix, that is, $\boldsymbol{D}_1 \propto \boldsymbol{I}$ ($\lambda_{11} = \lambda_{12} = \cdots = \lambda_{1K}$), as the next result shows, the design space can be reduced further.

Theorem 7.4.2. *For $\boldsymbol{D}_1 = c^{-1}\boldsymbol{I}$, the design vector $\boldsymbol{d}$, the constant c, the Maha-*

lanobis distance Δ *and the angle* $\phi_{z_1^+} = ang(\tilde{\boldsymbol{w}}_{z_1^+}, \tilde{\boldsymbol{\omega}}^*)$ *are sufficient to determine the power function* $\beta_{t_{AD}^+}$.

Proof. Let $\boldsymbol{R}$ be a rotation matrix with rotation axis $\tilde{\boldsymbol{\omega}}^*$ and define $\dot{\boldsymbol{w}}_{z_1^+} = \boldsymbol{R}\tilde{\boldsymbol{w}}_{z_1^+}$. Then note that we can write $\tilde{\boldsymbol{w}}_{z_j^+}$ in (7.5) as in (7.11) and $\boldsymbol{D}_j$ in (7.9) as $\boldsymbol{D}_j = \boldsymbol{R}^T\dot{\boldsymbol{D}}_j\boldsymbol{R}$, where

$$\dot{\boldsymbol{D}}_j = c^{-1}\boldsymbol{I} + (n_{(j-1)} - 1)\boldsymbol{S}_{\dot{\boldsymbol{w}}_{\boldsymbol{x}_{(j-1)}}} + \frac{n_0 n_{(j-1)}}{n_0 + n_{(j-1)}}\left(\dot{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j-1)}} - \dot{\boldsymbol{w}}_{z_1^+}\right)\left(\dot{\boldsymbol{w}}_{\bar{\boldsymbol{x}}_{(j-1)}} - \dot{\boldsymbol{w}}_{z_1^+}\right)^T.$$

Therefore, the weighting vector $\tilde{\boldsymbol{w}}_{t_j^+}$ in (7.7) can be written as

$$\tilde{\boldsymbol{w}}_{t_j^+}\boldsymbol{R}^T\dot{\boldsymbol{w}}_{t_j^+}, \qquad \dot{\boldsymbol{w}}_{t_j^+} = \dot{\boldsymbol{D}}_j^{-1}\dot{\boldsymbol{w}}_{z_j^+}$$

and hence the stage-wise test statistics, t_j^+ which, along with the design vector $\boldsymbol{d}$, fully describe the rejection region of the t_{AD}^+ test, can be re-written as in (7.12). Therefore, for $\boldsymbol{D}_1 = c^{-1}\boldsymbol{I}$, the rejection region of the t_{AD}^+ test after the rotation is equal to the initial but with $\tilde{\boldsymbol{w}}_{z_1^+}$ replaced by $\dot{\boldsymbol{w}}_{z_1^+}$ and $\tilde{\boldsymbol{w}}_{\boldsymbol{x}_{ij}}$ replaced by $\dot{\boldsymbol{w}}_{\boldsymbol{x}_{ij}}$. Importantly, $\dot{\boldsymbol{w}}_{\boldsymbol{x}_{ij}}$ have the same distribution, $N_K(\tilde{\boldsymbol{\omega}}^*, \boldsymbol{I})$, as $\tilde{\boldsymbol{w}}_{\boldsymbol{x}_{ij}}$, $i = 1, 2, \ldots, n_j$, $j = 1, 2, \ldots, J$. The result then follows using the same steps as in the proof of theorem 7.3.1. □

The last theorem proves that, for $\boldsymbol{D}_1 \propto \boldsymbol{I}$, we can use the fact that the prior $\boldsymbol{\Sigma}-$deviation matrix $\boldsymbol{D}_1$ does not change the directions of $\tilde{\boldsymbol{w}}_{z_j^+}$'s, to show that the relation of $\beta_{t_{AD}^+}$ to the model parameters and their prior estimates can be described simply by the scalars Δ and $\phi_{z_1^+} = ang(\tilde{\boldsymbol{w}}_{z_1^+}, \tilde{\boldsymbol{\omega}}^*)$.

7.5 Conclusions

The above results allows us to characterize the power function of linear combination tests and particularly its relation to the model parameters and the weighting vector. This power characterization provides a tool for understanding and alleviating

to some extent the complexities of multivariate tests and especially directional multivariate tests. The possibly high dimensional model parameters and the weighting vector/prior estimates are reduced to low dimensional summaries which are still sufficient to compute power. Importantly, these summaries have interpretations directly related to the strength of the treatment effect and the distance between the targeted direction and the effect direction. They provide a method for performing simple power analysis, but also understanding the behavior of linear combination tests.

The methods used to derive the power characterization are also interesting in their own right. They can be generally described by two steps: standardization and rotation invariance. The first standardization step is a prevalent technique for re-expressing statistical models in the standard deviation unit and eliminating correlations. Here, it allows us to re-express the weighting vector selection, which involves estimating the unknown model parameters, as a procedure of learning a single vector, that is, the optimal weighting vector, or, in other words, the effect direction. The second step of establishing a rotation invariance property for the power function allows us to identify the measure quantifying the angular distance between the selected and the optimal weighting vector, reducing further the design space. In the next chapter, we use these results, and particularly those in theorems 7.3.1, 7.4.1 and 7.4.2 to perform power analysis studies.

Chapter 8

Power analysis

8.1 Introduction

In this chapter, we study the power behavior of various multivariate global tests. We mainly focus on the tests developed in this thesis, that is, the z^+, t^+ and adaptive z^+_{AD}, t^+_{AD} linear combination tests, but we also examine the class of single-stage and group sequential linear combination z, t and z_{GS}, t_{GS} tests, with fixed weighting vectors, as described in sections 3.3.3 and 5.3.1 (pp. 37, 72), respectively. This includes, as special cases, O'Brien's OLS and GLS (for $\mathbf{\Sigma}$ known) tests in O'Brien [1984] and their group sequential extensions as described in Tang et al. [1993, 1989b]. Single-stage and group sequential χ^2, χ^2_{GS} and Hotelling's T^2, T^2_{GS} tests, as described in sections 3.3.1 and 5.3.1 (pp. 31, 72), are also examined, while Läuter [1996] exact SS and PC tests are considered for the applications in section 8.4.

In this study, we use the results of previous chapters, and especially chapter 7, which reduces substantially the complexities in performing power analysis of multivariate global tests and particularly directional tests. As we have seen in the previous chapter, for determining the power of linear combination tests, the model parameters and the weighting vector (or the prior estimates used to select it) can be reduced to two easily interpretable distance measures. The first is the Mahalanobis

distance, Δ, and expresses the strength of the treatment effect. The second is the distance of the selected weighting vector to the optimal weighting vector, $\tilde{\boldsymbol{\omega}}^*$. The latter can be seen as the distance of the "targeted" direction and the real effect direction. For the z, t, z_{GS} and t_{GS} tests, with fixed weighting vector, $\tilde{\boldsymbol{w}}$, this distance can be measured by the angle

$$\phi_{\boldsymbol{w}} = ang(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{\omega}}^*), \tag{8.1}$$

while for the z^+, z^+_{AD} with first stage weighting vector $\tilde{\boldsymbol{w}}_{z_1^+}$ using the angle

$$\phi_{z_1^+} = ang(\tilde{\boldsymbol{w}}_{z_1^+}, \tilde{\boldsymbol{\omega}}^*). \tag{8.2}$$

As we discuss later, this distance can also be studied in a simple way in the t^+ and t^+_{AD} tests.

These results provide a method to study complex effect structures, by looking at simple ones, provided that the latter keep the same values of the above summary measures as the former. For example, to study the power behavior of *any* linear combination test with weighting vector, $\tilde{\boldsymbol{w}}$, having angle, ϕ, to the optimal, $\tilde{\boldsymbol{\omega}}^*$, it is sufficient to set $\tilde{\boldsymbol{\omega}}^* = (1, 0, \ldots, 0)^T$ and $\tilde{\boldsymbol{w}} = (\cos\phi, \sin\phi, 0, \ldots, 0)^T$, as this structure sets $ang(\tilde{\boldsymbol{w}}, \tilde{\boldsymbol{\omega}}^*) = \phi$. Such results, which simplify power analysis considerably, are used in the empirical studies of the next sections.

For single-stage tests, the main focus of power analysis is on the study of two measures: power and sample size. For sequential or adaptive tests where early stopping is permitted, in addition to power and the total or maximum sample size, $n_T = n_1 + n_2 + \cdots + n_J$, it is useful to consider the sample size, n, actually used in the test. This can be smaller than n_T, due to early terminations, and depends on the observations' distribution which implies that it is a random variable. A common approach is then to study the total sample size and expected sample size, $E(n)$, for a fixed power level.

In the power analysis to follow, we consider another typical approach where the total sample size n_T is fixed at a certain value and the power of the test along with the expected sample size are examined. In particular, the expected sample size is studied using a relative, to the total sample n_T, measure, the rate of sample size reduction (RSSR). RSSR is defined as

$$RSSR = 100 \times \left(\frac{n_T - E(n)}{n_T} \right) \%. \tag{8.3}$$

The expected sample size $E(n)$ is approximated by the empirical mean of n across the simulated replications of the experiment. Similarly, the power of each test is approximated by the rate of rejection across the number of replications ($R = 50000$).

RSSR is linearly decreasing with $E(n)$ and thus larger values of $E(n)$ imply smaller values of RSSR. Furthermore, single-stage tests have $RSSR = 0$, as in single-stage tests no early stopping is permitted and thus $E(n) = n_T$. It is also useful to note that in our examples to follow, the two-stage z_{GS}, t_{GS} tests with fixed weighting vector, $\boldsymbol{w}$, have the same RSSR as the two-stage z^+_{AD}, t^+_{AD} tests (for the same design parameters), since without loss of generality we take $\boldsymbol{w} = \boldsymbol{w}_{z^+_1}$ or $\boldsymbol{w} = \boldsymbol{w}_{t^+_1}$ and thus these tests have equal number of early rejections.

To derive the critical values of the group sequential tests we use the gsDesign package in R software [Anderson, 2013]. In particular, for designs allowing for early acceptance ($\alpha_{j,0} < 1,\ j = 1, 2, \ldots, J-1$) we use the default Hwang-Shih-DeCani spending function approach in Hwang et al. [1990], while for designs not allowing for early acceptance ($\alpha_{j,0} = 1,\ j = 1, 2, \ldots, J-1$) the rejection critical values are chosen based on the methodology of Wang and Tsiatis [1987] as explained in Jennison and Turnbull [2000]. For the latter, in our presentations, the values of the constants C_{WT} and D_{WT} (as in Jennison and Turnbull [2000] notation) which are used to derive the Wang & Tsiatis stopping boundaries are provided along with the resulting critical values. Note that critical values derived from both Hwang et al.

[1990] and Wang and Tsiatis [1987] methodology control the type I error rate of group sequential z_{GS} tests. For group sequential t_{GS}, χ^2_{GS} and T^2_{GS} tests, we use the "significance level approach" as described in Jennison and Turnbull [2000]. For this, the $p-$values of the $j-$th stage-wise statistics of the latter tests are compared with the critical values $\alpha_{j,1} = 2(1 - \Phi(z_{\alpha_{j,1}}))$, where $z_{\alpha_{j,1}}$ the critical $z-$value for the group sequential z_{GS} test. As Jennison and Turnbull [2000] explains, these tests satisfy the type I error requirement only approximately, but investigations have shown that this approximation is remarkably accurate. Finally, to make the results of multi-stage tests more comparable, we set the first stage critical value of the adaptive z^+_{AD} and t^+_{AD} tests equal to the group sequential tests, while, to ensure type I error control, the second stage (or later stages) critical values of z^+_{AD} and t^+_{AD} are selected to satisfy the type I error equation in (6.10) (see page 89).

This chapter proceeds as follows. In section 8.2 we discuss the main outcomes of an extensive empirical study of the relation between the design and model parameters to the power behavior of the aforementioned global tests. In section 8.3, we provide a comparison of the power performance of these tests. In section 8.4, we consider the power performance of these global tests in the application to neuroimaging studies. We close the chapter by a brief discussion of the main results.

8.2 Design and model parameters

In the following, we discuss the main results of an extensive empirical study which attempts to describe the effect of design and model parameters on power behavior of various global tests. In each of the next sections, we mainly focus on the effect of individual design or model parameters on power and RSSR. We start with the sample size and global effect size.

Sample and global effect size

Generally, the power of well-behaved tests is expected to increase in the sample and effect size. As we have explained earlier, the Mahalanobis distance, Δ, is a measure of the size of the global effect. In particular, $\Delta = 0$ implies that the null hypothesis H_0 is true and thus the power of exact tests is expected to be equal to the significance level. As we can see in figure 8.1, for $\Delta = 0$ all the considered tests have power equal (at least approximately) to the significance level $\alpha = 0.05$. This confirms the analytical proofs of type I error control in these tests.

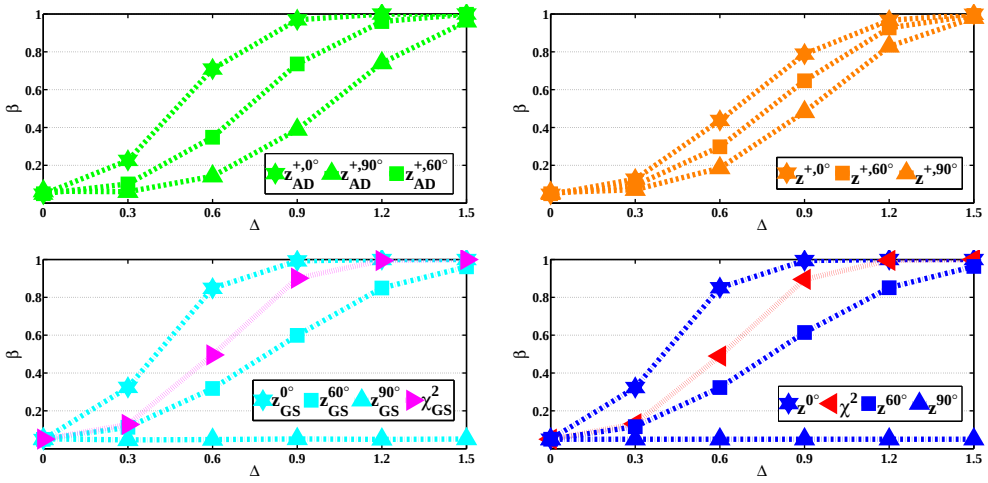


Figure 8.1: Power versus Mahalanobis distance Δ for various tests and angles (as indicated in the legend) of the weighting vectors $\tilde{\boldsymbol{w}}_{z_1^+} = \tilde{\boldsymbol{w}}$ to the optimal $\tilde{\boldsymbol{\omega}}^\star$. Here $\alpha = 0.05$, $\alpha_{0,1} = 1$, $\alpha_{1,1} = 0.0085$ ($D_{WT} = 0.1$, $C_{WT} = 1.994$), $K = 10$, $n_0 = 10$ and $n_T = 25$.

Larger values of Δ and/or n_T result in higher values of the location of the distribution of all the considered test statistics and thus, as we show in figures 8.1 and 8.2, higher power levels are attained. The exceptions are the z, t, z_{GS} and t_{GS} tests with fixed weighting vectors $\tilde{\boldsymbol{w}}$ orthogonal to $\tilde{\boldsymbol{\omega}}^+$ (see second row, figure 8.1 and 8.2) for which the location parameter is equal to 0 for any value of Δ and/or n_T and thus $\beta = \alpha$. On the contrary, the power of z^+ and t^+ and z^+_{AD}, t^+_{AD} tests increases with Δ and/or n_T even if $\tilde{\boldsymbol{w}}_{z_1^+}$ is orthogonal to the optimal $\tilde{\boldsymbol{\omega}}^+$ due to the

power gain after the interim adaptation of the weighting vector.

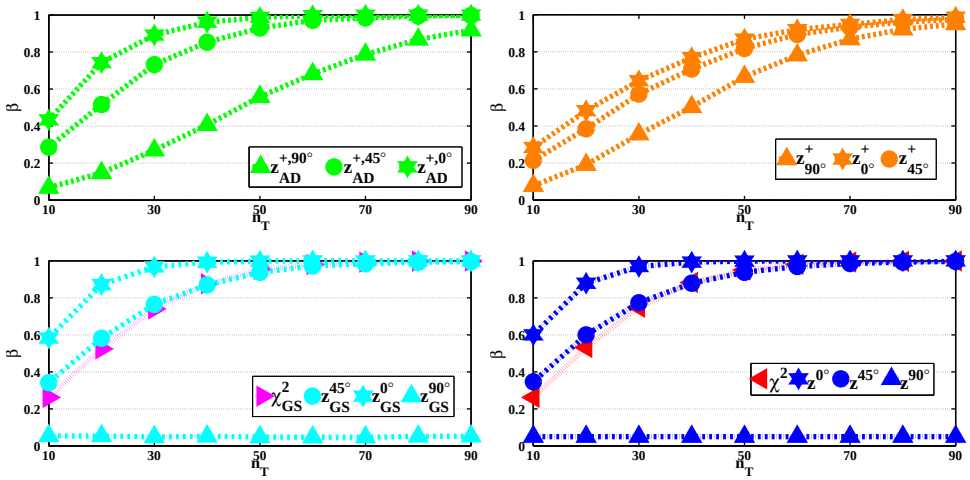


Figure 8.2: Power versus the total sample size n_T for various tests and angles (as indicated in the legend) of the weighting vectors $\tilde{\boldsymbol{w}}_{z_1^+} = \tilde{\boldsymbol{w}}$ to the optimal $\tilde{\boldsymbol{\omega}}^\star$. Here $\alpha = 0.05$, $\alpha_{0,1} = 1$, $\alpha_{1,1} = 0.0085$ ($D_{WT} = 0.1$, $C_{WT} = 1.994$), $K = 10$, $n_0 = 10$ and $\Delta = 0.7$.

Furthermore, larger Δ and/or n_T increase the chance to stop early and therefore they decrease the expected sample size of adaptive and sequential tests. This is illustrated by figure 8.3 which shows the increase in the rate of (total) sample size reduction, RSSR, for the z_{GS}/z^+_{AD} and χ^2_{GS}. Note that single-stage tests, as well as two-stage z and t tests with first stage weighting vector orthogonal to optimal, have $RSSR = 0$.

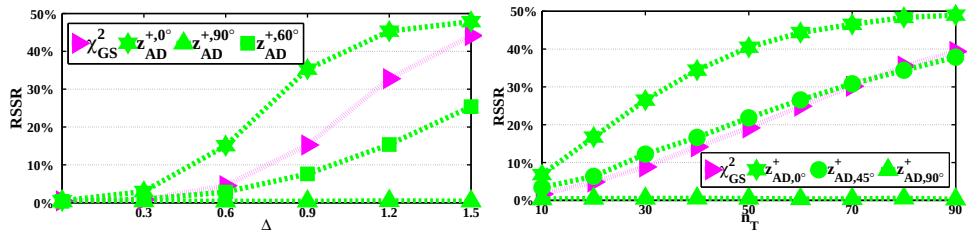


Figure 8.3: RSSR versus Mahalanobis distance Δ (left) and the total sample size n_T (right) for various tests and angles (as indicated in the legend) of the weighting vectors $\tilde{\boldsymbol{w}}_{z_1^+} = \tilde{\boldsymbol{w}}$ to the optimal $\tilde{\boldsymbol{\omega}}^\star$. Here $\alpha = 0.05$, $\alpha_{0,1} = 1$, $\alpha_{1,1} = 0.0085$ ($D_{WT} = 0.1$, $C_{WT} = 1.994$), $K = 10$, $n_0 = 10$ and $n_T = 25$ (left), $\Delta = 0.7$ (right).

Prior sample size

The prior sample size n_0 determines the weight of the prior estimates at the interim selections of the weighting vector. Larger values of n_0 enforce z^+, t^+ and z^+_{AD}, t^+_{AD} to have more similar power behavior to z, t and z_{GS} and t_{GS} tests with fixed weighting vector. This is because, for larger n_0, the first stage weighting vector, determined using the prior estimates, has larger weight in computing the weighting vector of later stages, while, for $n_0 \gg n_T$, the weighting vectors of all stages are approximately equal to the first stage weighting vector.

As we can see in figure 8.4, for $\boldsymbol{w}_{z_1^+}$ close to optimal, power is increasing with n_0, while for $\boldsymbol{w}_{z_1^+}$ far from optimal, power is decreasing with n_0. Moderate values of n_0, such as $n_0 = (0.75n_1, 1.25n_1)$, make the prior estimates influential, but do not dominate the accumulated data in selecting the weighting vector. This allows us to distinguish between z^+, t^+ and z^+_{AD}, t^+_{AD} test and the z, t and z_{GS} and t_{GS} tests.

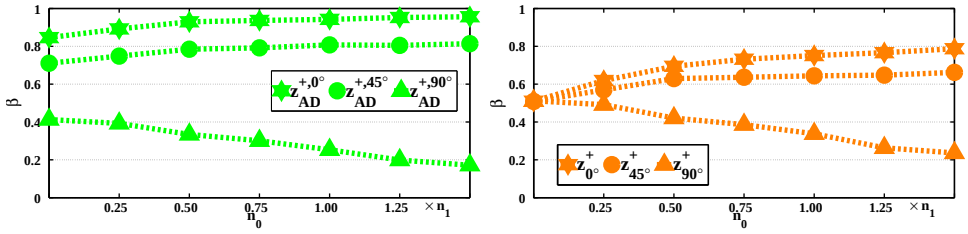


Figure 8.4: Power versus the prior sample size n_0 for z^+_{AD} (left) and z^+ (right) tests and angles (as indicated in the legend) of the weighting vectors $\tilde{\boldsymbol{w}}_{z_1^+}$ to the optimal $\tilde{\boldsymbol{\omega}}^\star$. Here $\alpha = 0.05$, $\alpha_{0,1} = 1$, $\alpha_{1,1} = 0.0085$ ($D_{WT} = 0.1$, $C_{WT} = 1.994$), $K = 10$, $n_T = 30$ and $\Delta = 0.75$.

Critical values

The rejection and acceptance critical values, $\alpha_{j,1}$, $\alpha_{j,0}$, $j = 1, 2, \ldots, J$, respectively, are chosen to satisfy a single constraint: the type I error equation of the test. This permits for substantial flexibility in selecting the critical values, which can be freely chosen to reflect, for instance, the investigators priorities on early stopping. In addition, as we confirm here, the critical values affect the power behavior of the

tests.

In general, larger values of the rejection critical values $\alpha_{j,1}$, $j = 1, 2, \ldots, J$, increase (decrease) power in cases where the potential power gain at the $j-$th stage is larger (smaller) compared to other stages. That is the case when n_j is larger (smaller) or when the $j-$th stage weighting vector is close to (far from) optimal. Generally, power is expected to be larger for greater $\alpha_{j,1}$ in late rather than early stages as in late stages greater samples are collected. On the other hand, larger values of $\alpha_{j,1}$ at early stages increase the chance to terminate the study early and thus the value of RSSR is expected to increase.

These arguments are supported by the results in table 8.1. Here the power and RSSR of two-stage χ^2_{GS} and z_{GS}, z^+_{AD} tests, the latter with weighting vectors of various angles to optimal, are presented. As we can see, power decreases with larger first stage critical values $\alpha_{1,1}$ since this result in smaller second stage critical values $\alpha_{2,1}$. On the contrary, larger $\alpha_{1,1}$ increases RSSR since it increases the chance for early stopping. Moderate values of $\alpha_{1,1}$ seem to achieve a good compromise between the benefits of high power and low expected sample size. It may be worth noting that, as $\alpha_{1,1}$ increases, the decrease in power and increase in RSSR is larger for more sample allocated to the first rather than the second stage, as in the former case the first stage study can potentially attain higher power.

Considering the futility critical values $\alpha_{j,0}$, $j = 1, 2, \ldots, J$, larger values increase the chance to proceed to (and thus reject at) subsequent stages, but, on the other hand, decrease the value of the rejection critical values due to the type I error restriction. In table 8.2, the power of non-single-stage tests is slightly changed with $\alpha_{1,0}$, with an exception being the z^+_{AD} test with the first stage weighting vector orthogonal to the optimal for which the power is substantially increased. This is because of the substantial power increase due to the second stage in which the weighting vector is no longer orthogonal, but possibly much closer, to the optimal. However, smaller $\alpha_{j,0}$ increase RSSR due to the larger chance of early acceptance.

Table 8.1: Power and RSSR versus the first-stage rejection critical value $\alpha_{1,1}$ for various two-stage ($J = 2$) tests and angles of the weighting vectors $\tilde{\boldsymbol{w}}_{z_1^+} = \tilde{\boldsymbol{w}}$ to the optimal $\tilde{\boldsymbol{\omega}}^\star$. Here $\alpha = 0.05$, $\Delta = 1$, $K = 10$, $n_T = 20$, $n_1 = 0.5n_T$, $n_0 = 5$.

		D_{WT}	0.00	0.10	0.25	0.40	0.50	0.75	1.00
		$\alpha_{1,1}$	0.0051	0.0086	0.0153	0.0237	0.0293	0.0419	0.0477
GS		$\alpha_{2,1}$	0.0477	0.0466	0.0414	0.0349	0.0293	0.0155	0.0051
χ^2_{GS}		β	0.89	0.89	0.88	0.87	0.86	0.81	0.74
		$RSSR$	11.75	14.41	18.10	21.06	22.70	25.58	26.84
z_{GS}	90°	β	0.05	0.05	0.05	0.05	0.05	0.05	0.05
		$RSSR$	0.26	0.44	0.82	1.18	1.48	2.14	2.51
	45°	β	0.88	0.88	0.87	0.86	0.85	0.81	0.74
		$RSSR$	14.41	17.49	21.23	24.47	26.17	29.03	30.01
	0°	β	0.99	0.99	0.99	0.99	0.99	0.98	0.97
		$RSSR$	32.17	35.42	38.42	40.77	41.96	43.45	44.01
AD		$\alpha_{2,1}$	0.0085	0.0087	0.0083	0.0070	0.0059	0.0026	0.0008
z^+_{AD}	90°	β	0.42	0.42	0.42	0.40	0.38	0.29	0.21
		$RSSR$	0.26	0.44	0.82	1.18	1.48	2.14	2.51
	45°	β	0.86	0.86	0.86	0.85	0.85	0.81	0.75
		$RSSR$	14.41	17.49	21.23	24.47	26.17	29.03	30.01
	0°	β	0.96	0.96	0.96	0.96	0.96	0.95	0.93
		$RSSR$	32.17	35.42	38.42	40.77	41.96	43.45	44.01

Table 8.2: Power and RSSR versus the first-stage futility critical value $\alpha_{1,0}$ for various two-stage ($J = 2$) tests and angles of the weighting vectors $\tilde{\boldsymbol{w}}_{z_1^+} = \tilde{\boldsymbol{w}}$ to the optimal $\tilde{\boldsymbol{\omega}}^\star$. Here $\alpha = 0.05$, $\alpha_{1,1} = 0.0148$, $n_0 = 7$, $K = 10$, $n_T = 20$, $n_1 = 0.5n_T$, $\Delta = 1$.

		$\alpha_{1,0}$	0.60	0.70	0.80	0.90	1.00
GS		$\alpha_{2,1}$	0.0448	0.0438	0.0432	0.0428	0.0420
z_{GS}	90°	β	0.05	0.05	0.05	0.05	0.05
		$RSSR$	35.5	33	30.5	28	0.05
	45°	β	0.87	0.87	0.87	0.87	0.88
		$RSSR$	23	22.5	22	22	21
	0°	β	0.99	0.99	0.99	0.99	0.99
		$RSSR$	38.5	38.5	38.5	38.5	38.5
AD		$\alpha_{2,1}$	0.0095	0.0091	0.0088	0.0086	0.0084
z_{AD}^+	90°	β	0.30	0.33	0.37	0.39	0.42
		$RSSR$	20.71	15.81	10.83	5.77	0.78
	45°	β	0.85	0.86	0.86	0.86	0.86
		$RSSR$	22.71	22.51	21.82	21.47	21.01
	0°	β	0.96	0.96	0.96	0.96	0.96
		$RSSR$	38.47	38.29	38.12	38.45	38.25

These results indicate that investigators should relate the selection of $\alpha_{j,0}$ to the degree of interest on early acceptance of H_0 as opposed to final rejection. It is worth noting here that even if $\alpha_{j,0}$, $j = 1, 2, \ldots, J$ are all set equal to 1, early stopping of the study either for futility or for other reasons (such as safety), without declaring rejection of H_0, does not inflate the type I error. Suitable thresholds for futility early stopping can thus be specified, but not taken into account for computing the type I error. Such procedures are often called non-binding stopping rules. Bretz et al. [2009] and Jennison and Turnbull [2000] provide methodologies and discussions on the subject.

Sample allocation

For group sequential and adaptive tests, an important issue is how to allocate the total sample size, n_T, to the different stages. As we have discussed in chapter 5, adaptive designs as well as some group sequential designs allow for sample size adaptation and thus n_T is allowed to change based on the interim results. However, in this section, to study the effect of different sample allocations on the power of global tests, we assume that n_T is fixed and we consider only two-stage designs. The sample allocation is described by $r_1 = n_1/n_T$.

Table 8.3 shows that the changes in the power of z_{GS} and χ^2_{GS} tests for different values of r_1 are negligible. This is partly due to the moderate value of $\alpha_{1,1}$. For very unbalanced $\alpha-$allocation to stages, for example $\alpha_{1,1} \lessapprox \alpha$, allocating sample size accordingly provides higher power. On the other hand, table 8.3 shows that equal sample allocation to stages ($r_1 = 0.5$) can give substantially larger RSSR.

The power of the z^+ test is greatly affected by the sample allocation, with small to moderate allocation ratios attaining substantially greater power levels. This is because z^+ allocates the α rate exclusively to the second stage. In table 8.4, higher power is observed for $r_1 = 0.25$, with power slightly lower for $r_1 = 0.5$ and substantially lower for $r_1 = 0.75$.

Table 8.3: Power and RSSR versus the sample allocation ratio $r_1 = n_1/n_T$ for the two-stage ($J = 2$) χ^2_{GS} and z_{GS} tests, the latter with different angles of the weighting vector $\tilde{\boldsymbol{w}}$ to the optimal $\tilde{\boldsymbol{\omega}}^\star$. Here $\alpha = 0.05$, $\alpha_{1,1} = 0.0153$ ($D_{WT} = 0.388$, 0.25, -0.175), $K = 10$, $\Delta = 0.85$, $n_T = 30$.

			$r_1 = 0.25$	$r_1 = 0.5$	$r_1 = 0.75$
χ^2_{GS}		β	0.91	0.91	0.91
		$RSSR$	12.14	20.06	15.45
z_{GS}	90°	β	0.05	0.05	0.05
		$RSSR$	1.17	0.79	0.37
	45°	β	0.90	0.90	0.90
		$RSSR$	17.09	23.07	15.81
	0°	β	1.00	1.00	1.00
		$RSSR$	36.09	40.36	22.16

In contrast to z^+ and t^+, the adaptive z^+_{AD} and t^+_{AD} tests allocate some of the α rate to the first stage, which greatly affect their power behavior. If the first stage weighting vector is very close to optimal, an increase in distance to optimality may occur at later stages due to sampling error, and thus, in contrast to z^+ and t^+, allocating greater sample sizes to the first stage may result in slightly greater power values. On the other hand, similarly to z^+ and t^+ tests, for first stage weighting vectors far from the optimal, allocating larger sample to the second stage is more appropriate as the first stage sample moves the second stage weighting vector closer to optimal and thus the potential power at the second stage is greater.

The results in table 8.4 show that the highest values of power of the z^+_{AD} test for angles between the first stage and the optimal weighting vector, 0°, 45° and 90°,

are respectively attained for $r_1 = 0.75$, 0.75 and 0.25.

Table 8.4: Power and RSSR versus the sample allocation ratio $r_1 = n_1/n_T$ for the z^+ and z^+_{AD} tests with different angles of the prior weighting vector $\tilde{\boldsymbol{w}}_{z_1^+}$ to the optimal $\tilde{\boldsymbol{w}}$. Here $\alpha = 0.05$, $\alpha_{1,1} = 0.0153$ ($D_{WT} = 0.388$, 0.25, -0.175), $K = 10$, $\Delta = 0.85$, $n_T = 30$. Note that for the z^+ tests RSSR$= 0$ and that for the two-stage z^+_{AD} test RSSR is not affected by n_0.

				$r_1 = 0.25$	$r_1 = 0.5$	$r_1 = 0.75$
z^+	$n_0 = 0.25n_1$	90°	β	0.64	0.64	0.42
		45°	β	0.74	0.72	0.47
		0°	β	0.77	0.75	0.49
	$n_0 = 0.75n_1$	90°	β	0.57	0.52	0.32
		45°	β	0.83	0.77	0.49
		0°	β	0.90	0.83	0.55
z^+_{AD}	$n_0 = 0.25n_1$	90°	β	0.55	0.53	0.32
		45°	β	0.77	0.87	0.88
		0°	β	0.86	0.96	0.98
	$n_0 = 0.75n_1$	90°	β	0.47	0.41	0.24
		45°	β	0.85	0.90	0.89
		0°	β	0.94	0.98	0.99
		90°	$RSSR$	1.14	0.71	0.35
		45°	$RSSR$	17.44	23.02	15.77
		0°	$RSSR$	36.23	40.21	22.20

However, most importantly, the loss in power between $r_1 = 0.5$ and the more extreme allocations, either greater or smaller r_1, is small, which makes equally sized stages overall more appropriate. The latter result is also supported by the results for RSSR, as the largest values of RSSR are attained for $r_1 = 0.5$ (except for angle 90° where RSSR is negligible).

It might be useful to note here that, overall, balanced allocations seem to be appropriate in most cases. The main exceptions are the situations where the allocation of the α rate to the different stages is very imbalanced (e.g. z^+ and t^+ tests), in which case it might be preferrable to allocate the sample accordingly.

8.2.1 Impact of $\boldsymbol{\Sigma}$ unknown

In this section, we consider the impact of $\boldsymbol{\Sigma}$ being unknown, and thus estimated, on the performance of the t^+_{AD} test. As we have discussed in section 7.4, the location parameters of the stage-wise t^+_j statistics are

$$\theta_{t^+_j} = \frac{\tilde{\boldsymbol{w}}^T_{z^+_j} \boldsymbol{D}_j^{-1} \tilde{\boldsymbol{\omega}}^*}{\|\boldsymbol{D}_j^{-1} \tilde{\boldsymbol{w}}_{z^+_j}\|} = \frac{\boldsymbol{c}^T_{z^+_j} \boldsymbol{\Lambda}_j^{-1} \boldsymbol{c}^*}{\|\boldsymbol{\Lambda}_j^{-1} \boldsymbol{c}_{z^+_j}\|}, \qquad j = 1, 2, \ldots, J,$$

where $\boldsymbol{D}_j = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{S}_{j-1} \boldsymbol{\Sigma}^{-1/2}$, $\boldsymbol{\Lambda}_j = Diag(\boldsymbol{\lambda}_j)$ the $\boldsymbol{\Sigma}-$deviation matrix, with $\boldsymbol{\lambda}_j$ being the vector of eigenvalues of $\boldsymbol{D}_j$ and the vectors $\boldsymbol{c}_{z^+_j} = \boldsymbol{V}^T \boldsymbol{w}_{z^+_j}$, $\boldsymbol{c}^* = \boldsymbol{V}^T \tilde{\boldsymbol{\omega}}^*$, where $\boldsymbol{V}$ is the matrix of eigenvectors of $\boldsymbol{D}_1$.

We firstly consider the situation where $\boldsymbol{D}_1 = \boldsymbol{\Sigma}^{-1/2} \boldsymbol{S}_0 \boldsymbol{\Sigma}^{-1/2}$ is proportional to the identity matrix $\boldsymbol{I}_K$ ($\boldsymbol{\lambda}_1 \propto \boldsymbol{1}_K$). This situation arises if the prior estimate $\boldsymbol{S}_0$ is proportional to $\boldsymbol{\Sigma}$. For $\boldsymbol{D}_1 \propto \boldsymbol{I}_K$, as we showed in theorem 7.4.2, the power of t^+_{AD} can be studied in a similar way to z^+_{AD}, that is based on Δ and the angle between the vector $\tilde{\boldsymbol{w}}_{t^+_1} = \tilde{\boldsymbol{w}}_{z^+_1}$ and the optimal $\tilde{\boldsymbol{\omega}}^*$.

It is worth considering the difference in power behavior of t^+_{AD} between this simpler case and the case where $\boldsymbol{D}_1 \not\propto \boldsymbol{I}_K$. In the former case, the direction of $\tilde{\boldsymbol{w}}_{t^+_j}$ is mostly influenced by $\tilde{\boldsymbol{w}}_{z^+_j}$ and less by $\boldsymbol{D}_j$. For the first stage, the weighting

vector $\tilde{\boldsymbol{w}}_{t_1^+} = \tilde{\boldsymbol{w}}_{z_1^+}$ while for subsequent stages $\boldsymbol{D}_j$ are expected to be closer to $\boldsymbol{I}_K$ and thus $\tilde{\boldsymbol{w}}_{t_j^+}$ closer to $\tilde{\boldsymbol{w}}_{z_j^+}$. In other words, having a precise (up to a constant) prior covariance matrix estimate reduces the estimation variability and we generally expect $\tilde{\boldsymbol{w}}_{t_j^+}$ to be closer to $\tilde{\boldsymbol{w}}_{z_j^+}$.

If $\boldsymbol{D}_1 \not\propto \boldsymbol{I}_K$ ($\boldsymbol{\lambda}_1 \not\propto \mathbf{1}_K$), the direction of $\boldsymbol{\lambda}_1$ is more influential on $\tilde{\boldsymbol{w}}_{t_j^+}$. To illustrate this situation with an example, consider the case where the covariance matrix and its prior estimate $\boldsymbol{S}_0$ are diagonal, that is $\boldsymbol{\Sigma} = Diag\left(\sigma_1^2, \ldots, \sigma_K^2\right)$, $\boldsymbol{S}_0 = Diag\left(s_{01}^2, \ldots, s_{0K}^2\right)$. Hence, the matrix $\boldsymbol{D}_1^{-1} = Diag\left((\sigma_1/s_{01})^2, \ldots, (\sigma_K/s_{0K})^2\right)$ which implies that variance underestimation of say the $l-$th variable results in $\sigma_l/s_{0l} > 1$. In this case, $\boldsymbol{D}_1^{-1}$ gives greater weight to the $l-$th variable, compared to $\boldsymbol{D}_1 \propto \boldsymbol{I}_K$, which agrees with the intuition that variables which are expected to be less variant should receive more weight. Overestimation of the variance, on the other hand, leads to $\sigma_l/s_l < 1$ and thus smaller $l-$weight, which also seems appropriate intuitively, since variables expected to be more variant should receive less weight.

This example illustrates that the consequences of $\boldsymbol{D}_1 \not\propto \boldsymbol{I}_K$ or equivalently $\boldsymbol{\lambda}_1 \not\propto \mathbf{1}_K$ on power are double-edged. That is, compared to the situation of $\boldsymbol{\lambda}_1 \propto \mathbf{1}_K$, the distance of $\tilde{\boldsymbol{w}}_{t_j^+}$ to optimal can be larger but also smaller depending on how close the direction of $\boldsymbol{\lambda}_1$ is to the optimal direction $\boldsymbol{c}^*$. The former situation arises in situations where, compared with $\boldsymbol{\lambda}_1 \propto \mathbf{1}_K$, the prior estimate $\boldsymbol{S}_0$ gets $\tilde{\boldsymbol{w}}_{z_j^+}$ more distant to the optimal (for instance variance underestimation of a variable which should receive less weight) and the latter in situations where $\boldsymbol{S}_0$ brings $\tilde{\boldsymbol{w}}_{z_j^+}$ closer to optimal (for instance variance underestimation of a variable which should receive more weight).

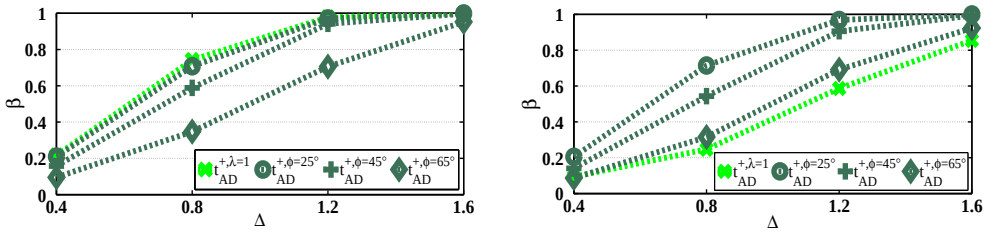


Figure 8.5: Power of the t^{+}_{AD} test versus Mahalanobis distance for various $\boldsymbol{c}^{\star}, \boldsymbol{c}_{z_1^+}, \boldsymbol{\lambda}_1$. In the left panel, the vectors $\boldsymbol{c}^{\star} = \boldsymbol{c}_{z_1^+} \propto \mathbf{1}_K$ while in the right panel $\boldsymbol{c}^{\star} = \boldsymbol{e}_1 = (1, 0, \ldots, 0)^T$ and $\boldsymbol{c}_{z_1^+} \propto \mathbf{1}_K$ which, for $\boldsymbol{\lambda}_1 = \mathbf{1}_K$, give $\phi = ang(\boldsymbol{c}^{\star}, \boldsymbol{\Lambda}_1^{-1}\boldsymbol{c}_{z_1^+}) = ang(\boldsymbol{c}^{\star}, \boldsymbol{\lambda}_1^{-1}) = 0°$ and $72°$, respectively. In both panels, $\boldsymbol{\lambda}_1 \not\propto \mathbf{1}_K$ are also chosen to give $\phi = 25°$, $45°$ and $65°$ (as indicated in the legend). The remaining design parameters are $J = 2$, $K = 10$, $\alpha = 0.05$, $\alpha_{1,1} = 0.01$, $\alpha_{0,1} = 1$, $n_T = 20$, $r_1 = 0.5$, $n_0 = 0.75 n_1$, $\nu_0 = n_0 - 1$.

Both situations are illustrated by respectively the left and right panel in figure 8.5. In the left panel, $\boldsymbol{\lambda}_1 \propto \mathbf{1}_K$ has the same direction with $\boldsymbol{c}^{*} \propto \mathbf{1}_K$ and thus gives higher power than more distant $\boldsymbol{\lambda}_1$ to $\boldsymbol{c}^{*}$, while in the right panel $\boldsymbol{\lambda}_1 \propto \mathbf{1}_K$ has direction distant to the direction of $\boldsymbol{c}^{*}$ and thus $\boldsymbol{\lambda}_1$ closer to $\boldsymbol{c}^{*}$ can attain substantially greater power.

It is useful to note that throughout our simulations of t^{+}_{AD} test, the angle $\phi_{t_1^+} = ang(\tilde{\boldsymbol{\omega}}^{+}, \tilde{\boldsymbol{w}}_{t_1^+})$ is proved to be a robust summary of the angular distance between the model parameters and their prior estimates. This is not very surprising because it is a sufficient summary for the first stage power and a good indicator, albeit not sufficient (see figure 8.6), for the power of subsequent stages.

For the above reasons, but also to reduce complexity, in the comparisons to follow, we focus on the case of $\boldsymbol{\lambda}_1 \propto \mathbf{1}_K$ for various values of $\phi_{t_1^+}$. Note that, as we explain later on, in the simulations presented in the next section the case of $\boldsymbol{\lambda}_1 \propto \mathbf{1}_K$ resembles the case on the right panel of figure 8.5.

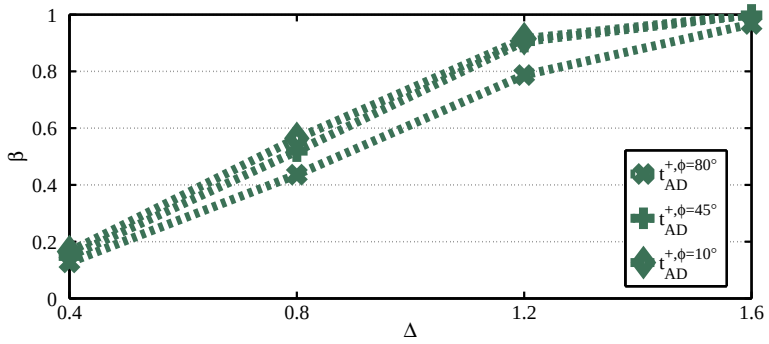


Figure 8.6: Power of the t^+_{AD} test versus Mahalanobis distance for various $\boldsymbol{c}^\star, \boldsymbol{c}_{z_1^+}, \boldsymbol{\lambda}_1$. The vector $\boldsymbol{c}^\star = \boldsymbol{e}_1$ and $\boldsymbol{c}_{z_1^+}$ is chosen to give $\phi = ang(\boldsymbol{c}^\star, \boldsymbol{c}_{z_1^+}) = 10°$, $45°$ and $80°$ (as indicated in the legend), while for all values of ϕ, $\boldsymbol{\lambda}_1$ is chosen to give the same $\phi_{t_1^+} = ang(\boldsymbol{c}^\star, \boldsymbol{\Lambda}_1^{-1}\boldsymbol{c}_{z_1^+}) = 45°$. The remaining design parameters are $J = 2$, $K = 10$, $\alpha = 0.05$, $\alpha_{1,1} = 0.01$, $\alpha_{0,1} = 1$, $n_T = 20$, $r_1 = 0.5$ and $n_0 = 0.75n_1$, $\nu_0 = n_0 - 1$.

8.3 Comparisons

In this section, we compare various global tests in terms of power performance. We first compare the tests for $\boldsymbol{\Sigma}$ known using various illustrations.

In figure 8.7, we present the single-stage z, z^+ and χ^2 tests. Here, the prior sample size $n_0 = 10$ ($n_0 \approx 0.75n_1$) which allows the prior estimates to be influential but not dominating of the pilot estimates. The z^+ loses power compared to the z test for small angles $\phi_{\boldsymbol{w}} = \phi_{z_1^+}$ (as in (8.1) and (8.2) page 119), mainly due to the use of the first stage sample only for selecting the weighting vector. On the other hand, the use of the pilot data to modify the weighting vector at the interim analysis results in a substantial power gain for $\phi_{\boldsymbol{w}} = \phi_{z_1^+}$ close to $90°$. Importantly, the z^+ test, in contrast to z, can achieve high power levels for sufficient values of n_T and/or Δ, even if $\tilde{\boldsymbol{w}}_{z_1^+}$ is orthogonal to optimal. This property of z^+ acts as a benchmark for the case the unknown direction of $\tilde{\boldsymbol{\omega}}^*$ is unexpectedly distant to the anticipated direction. The χ^2 test attains power around the levels of z^+ for $\phi_{z_1^+} = 45°$.

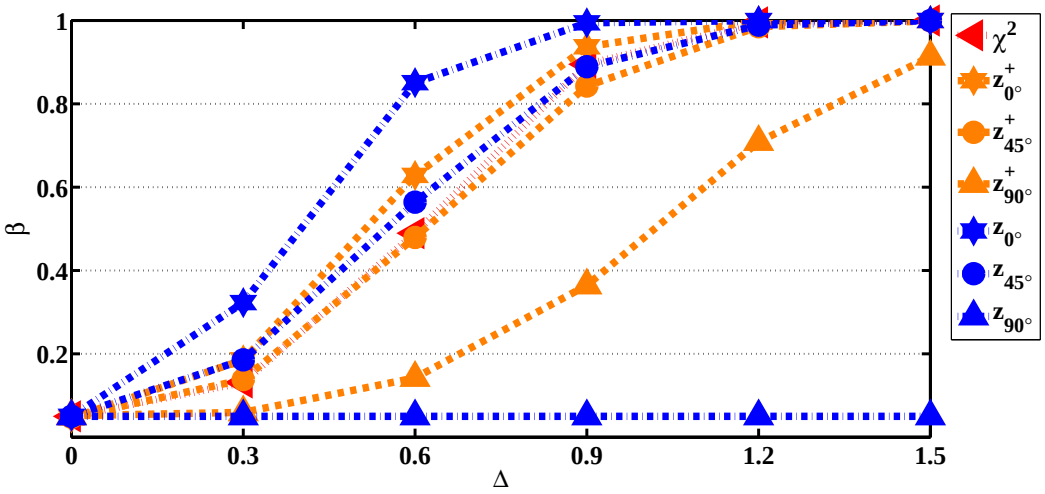


Figure 8.7: Power versus Mahalanobis distance Δ for the χ^2, z and z^+ tests, the latter with various angles (as indicated) of the weighting vectors $\tilde{\boldsymbol{w}}_{z_1^+} = \tilde{\boldsymbol{w}}$ to the optimal. Here $\alpha = 0.05$, $\alpha_{0,1} = 1$, $\alpha_{1,1} = 0.0085$ ($D_{WT} = 0.1$, $C_{WT} = 1.994$), $n_0 = 10$, $r_1 = 0.25$ and $n_T = 25$.

The loss of power in z^+ for small angles to the optimal, due to the zero α spending to the first stage, is reduced with the adaptive z^+_{AD} test. On the other hand, for very large angles, the adaptive z^+_{AD} loses small levels of power compared to z^+, due to the spending of α rate to the first stage where little power is attained. The differences between z^+_{AD} and z^+ are smaller for $r = 0.25$ (see first row figure 8.8), but they increase considerably for $r_1 = 0.5$ (see second row figure 8.8), as the amount of sample allocated to the first stage is larger. Further, the z^+_{AD} permits for early stopping and thus it has lower expected sample size than the total sample size used (by construction) in z^+.

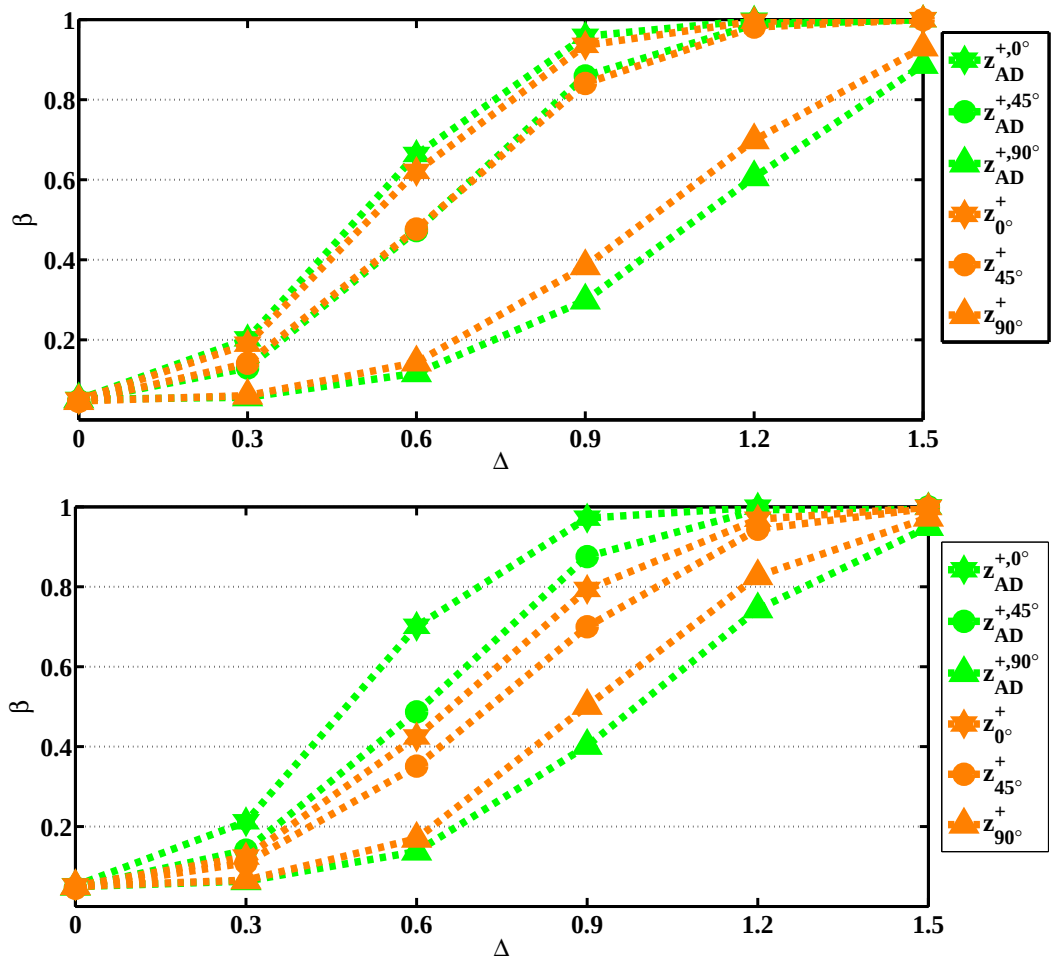


Figure 8.8: Power versus Mahalanobis distance Δ for z^+ and z^+_{AD} tests and angles (as indicated in the legend) of the weighting vectors $\tilde{\boldsymbol{w}}_{z_1^+}$. Here $\alpha = 0.05$, $\alpha_{0,1} = 1$, $\alpha_{1,1} = 0.0085$ ($D_{WT} = 0.1$, $C_{WT} = 1.994$), $n_0 = 10$, $r_1 = 0.25$ (first row), $r_1 = 0.25$ (second row) and $n_T = 25$.

Similarly to the single-stage tests, as shown in figure 8.9, the adaptive z^+_{AD}, compared to z_{GS}, loses power for small angles $\phi_{\boldsymbol{w}} = \phi_{z_1^+}$ of the first stage weighting vector to the optimal. This is due to the adaptation of the weighting vector, which may increase slightly the angle to optimality, due to sampling error, but also due to the use of adaptive combination tests which are less efficient than the group sequential tests. On the other hand, z^+_{AD} attains large power levels for any first weighting vector, even if the latter is orthogonal to the optimal, due to the interim adaptation

which can correct incorrect prior estimates. This is a substantial benchmark to ensure that, even in the worst case scenario, sufficient power can be attained in any situation for large enough effect and/or sample sizes.

The group sequential χ^2_{GS} tests is approximately equal to the power of the z^+_{AD} test for $\phi_{z_1^+}$ around 45°. Furthermore, as we can see in table 8.5, the χ^2_{GS} test has RSSR also close to the RSSR values of z_{GS}/z^+_{AD} tests with $\phi_{\boldsymbol{w}} = \phi_{z_1^+} = 45°$.

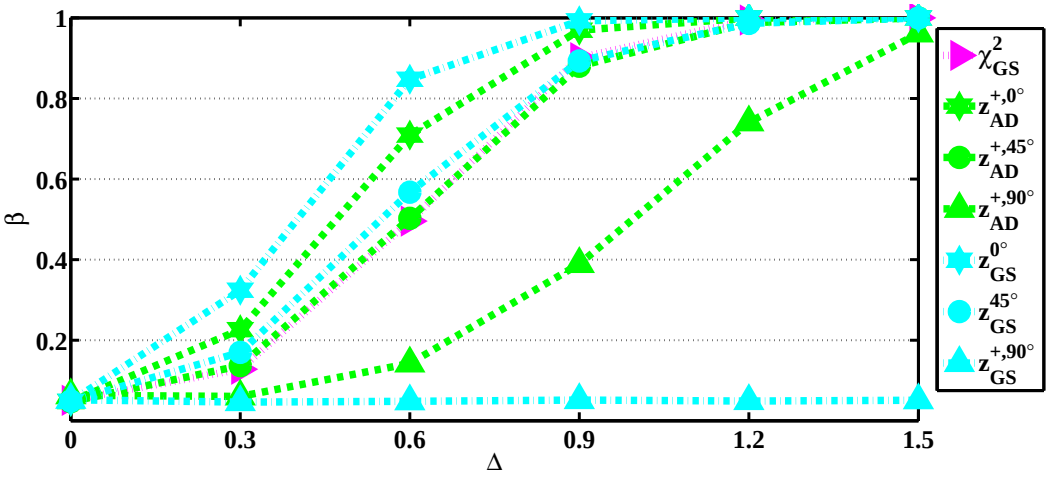


Figure 8.9: Power versus Mahalanobis distance Δ for various tests and angles of the weighting vectors $\tilde{\boldsymbol{w}}_{z_1^+} = \tilde{\boldsymbol{w}}$. Here $\alpha = 0.05$, $\alpha_{0,1} = 1$, $\alpha_{1,1} = 0.0085$ ($D_{WT} = 0.1$, $C_{WT} = 1.994$), $n_0 = 10$, $r_1 = 0.5$ and $n_T = 25$.

The fully multivariate χ^2 and χ^2_{GS} heavily rely on the assumption that $\boldsymbol{\Sigma}$ is known. For $\boldsymbol{\Sigma}$ unknown, the counterparts of χ^2 and χ^2_{GS}, Hotelling's T^2 and T^2_{GS} tests are greatly affected by the lack of knowledge of $\boldsymbol{\Sigma}$, especially for small sample sizes. As we can see in table 8.6, for $n_T \leq K = 10$ or n_T slightly larger than K, T^2 is respectively inapplicable or very inefficient with power levels lower than the power of t^+ even for angles close to orthogonal.

Table 8.5: RSSR versus Mahalanobis distance Δ for the χ^2_{GS} and the linear combination z_{GS} and z^+_{AD} tests, the latter with various angles of the weighting vectors $\tilde{\boldsymbol{w}}_{z_1^+} = \tilde{\boldsymbol{w}}$. Here $\alpha = 0.05$, $\alpha_{0,1} = 1$, $\alpha_{1,1} = 0.0085$ ($D_{WT} = 0.1$, $C_{WT} = 1.994$), $K = 10$, $n_0 = 10$, $r_1 = 0.5$ and $n_T = 25$.

	Δ	0	0.3	0.6	0.9	1.2	1.5
χ^2_{GS}		0.45	0.94	4.38	15.31	31.90	44.61
z_{GS},z^+_{AD}	90°	0.46	0.38	0.48	0.49	0.44	0.27
	45°	0.50	1.46	6.83	17.57	32.16	42.16
	0°	0.45	3.01	15.87	34.66	46.00	47.84

Table 8.6: Power versus total sample size n_T for the T^2_{GS} and the linear combination t_{GS} and t^+_{AD} tests, the latter with various angles of the weighting vectors $\tilde{\boldsymbol{w}}_{z_1^+} = \tilde{\boldsymbol{w}}$ to the optimal. Here $\alpha = 0.05$, $\alpha_{0,1} = 1$, $\alpha_{1,1} = 0.0154$ ($D_{WT} = 0.25$, $C_{WT} = 2.038$), $r_1 = 0.25$ and $\Delta = 1$.

	n_T	10	15	20	25	30	60
T^2		-	0.25	0.53	0.75	0.89	1.00
t	90°	0.05	0.05	0.05	0.05	0.05	0.05
	45°	0.51	0.72	0.85	0.92	0.96	1.00
	0°	0.80	0.95	0.99	1.00	1.00	1.00
t^+	90°	0.11	0.19	0.28	0.37	0.48	0.86
	45°	0.36	0.56	0.72	0.84	0.90	1.00
	0°	0.51	0.74	0.87	0.95	0.97	1.00

The group sequential T^2_{GS} is not applicable even for $2n_T < K$, while, as we can see in table 8.7, it has power close to the power of linear combination tests of

large angles (larger than 45°), even for $n_T = 30$. As sample size becomes considerably bigger than K, the power of T^2 and T^2_{GS} test increases sharply to yield power levels analogous to the χ^2 and χ^2_{GS} tests, that is β and $RSSR$ close to t^+/t^+_{AD} with $\phi_{t^+_1} = 45°$.

Table 8.7: Power versus total sample size n_T for the T^2_{GS} and the linear combination t_{GS} and t^+_{AD} tests, the latter with various angles of the weighting vectors $\tilde{\boldsymbol{w}}_{z^+_1} = \tilde{\boldsymbol{w}}$ to the optimal. Here $\alpha = 0.05$, $\alpha_{0,1} = 1$, $\alpha_{1,1} = 0.0154$ ($D_{WT} = 0.25$, $C_{WT} = 2.038$), $r_1 = 0.5$ and $\Delta = 1$.

		n_T	10	15	20	25	30	60
T^2_{GS}			-	-	-	0.73	0.88	1.00
t_{GS}		90°	0.05	0.05	0.05	0.05	0.05	0.05
		45°	0.49	0.70	0.84	0.91	0.96	1.00
		0°	0.79	0.94	0.99	1.00	1.00	1.00
t^+_{AD}	$n_0 = 1.25n_1$	90°	0.08	0.12	0.19	0.26	0.33	0.73
		45°	0.34	0.59	0.77	0.89	0.95	1.00
		0°	0.55	0.81	0.93	0.98	0.99	1.00

Considering the t^+ and t^+_{AD} test, it is useful to note here that in the results presented in this section we consider the case where the $\boldsymbol{\Sigma}-$deviation matrix $\boldsymbol{D}_1 = \boldsymbol{I}$ which, using the results of theorem 7.4.2 (see page 115), it can be studied in a similar way to the case of known $\boldsymbol{\Sigma}$. That is, based on Δ and the angle between the selected and the optimal weighting vector, $\phi_{t^+_1} = \phi_{z^+_1}$. For the results in tables 8.6, 8.7 and 8.8, the case of $\boldsymbol{D}_1 = \boldsymbol{I}$ can be thought of as representative of $\boldsymbol{\lambda}_1^{-1}$ fairly distant to $\boldsymbol{c}^*$ (right panel of figure 8.5), since we take $\boldsymbol{c}^* = \boldsymbol{e}_1$ resulting in $ang(\boldsymbol{c}^*, \boldsymbol{\lambda}_1^{-1}) = 71°$, for $K = 10$.

In terms of power performance, first note that the power of the linear combination t, t^+ and t^+_{AD} tests are also lower than their counterparts z tests (same design parameters), but not to the extend of the T^2 tests. Moreover, the patterns of power difference across these tests remain the same. That is, the t^+ and t^+_{AD} tests have lower power than t and t_{GS} tests for small angles of the weighting vector to the optimal, but gain substantial amounts of power for large angles.

Finally, as we can see in table 8.8, for small sample sizes, the T^2_{GS}, if applicable, attains similar values of RSSR to linear combination tests with weighting vectors nearly orthogonal to the optimal, but as the sample size becomes substantially larger than K, the values of RSSR of the T^2_{GS} are similar to those of linear combination tests with much smaller angles to the optimal. Specifically, analogous results to the χ^2 tests are attained.

Table 8.8: RSSR versus total sample size n_T for the T^2_{GS} and the linear combination t_{GS} and t^+_{AD} tests, the latter with various angles of the weighting vectors $\tilde{\boldsymbol{w}}_{z_1^+} = \tilde{\boldsymbol{w}}$ to the optimal. Here $\alpha = 0.05$, $\alpha_{0,1} = 1$, $\alpha_{1,1} = 0.0154$ ($D_{WT} = 0.25$, $C_{WT} = 2.038$), $r_1 = 0.5$ and $\Delta = 1$.

	n_T	10	15	20	25	30	60
T^2_{GS}		-	-	-	2.33	4.96	36.53
t_{GS},t^+_{AD}	90°	0.72	0.65	0.81	0.70	0.83	0.78
	45°	4.50	9.52	14.81	20.52	25.59	44.76
	0°	9.00	20.07	29.40	36.87	42.49	49.85

8.4 Application to neuroimaging studies

In this section, we consider application of various global tests to the neuroimaging studies introduced in sections 2.3.2 and 2.4.2 (see pp. 15, 20). As we have discussed earlier, the investigators of both the fMRI and EEG study were interested in evalu-

ating the null hypothesis of no treatment effect across various locations of the brain. The tests that we consider here are suitable for the assessment of this global null hypothesis. Specifically, we apply the Bonferroni global test, Hotelling's T^2 test, O'Brien's OLS t test, Läuter SS and PC t tests and the t^+ and t^+_{AD} tests.

Our main target is to assess the performance of these tests in the settings of these applications. For this, we set the design parameters, α and n, equal to those of each study and the model parameters, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, equal to their sample estimates extracted from the data of each study. It might be useful to note here that these studies are performed under a single-stage design. Thus, we set the total sample size of adaptive tests equal to the sample size of the study, while suitable allocations of the sample and α rate to the different stages are selected as described in the next sections. The power of the above tests is derived by simulated replications of the experiment under the selected parameter values.

8.4.1 Application to the fMRI study

The fMRI study as introduced in section 2.3.2 (page 15) was conducted for drug development purposes using $n_T = 13$ subjects.

As we have seen in table 2.2 (see page 17), effect sizes differ across ROI and generally high correlations are observed. The local effects are relatively small, resulting in low power $\beta_{BON} = 0.19$ for the Bonferroni global test. Despite the small local effects, the Mahalanobis distance $\Delta = 2.7657$ is relatively large suggesting strong global effects. Hotelling's T^2 test has larger, but still relatively low, power, $\beta_{T^2} = 0.40$, which can be attributed to the sample size being very close to the number of ROI.

The power of the optimal single-stage linear combination t test has power $\beta_{t^*} = 1.00$. This suggests that a linear combination test, where the search for effects is targeted in a specific direction, rather than the whole $11-$dimensional space as in T^2, can attain large power levels. O'Brien's OLS t test and Läuter's SS and PC

tests, which target on the uniform direction $\mathbf{1}_K$, have very low power $\beta_{OLS} = 0.14$, $\beta_{SS} = 0.13$, $\beta_{PC} = 0.14$. The reason is that the uniform direction is very distant to the optimal with angle $\phi_{\boldsymbol{w}} = 84.63°$. In other words, the low power values are a result of searching for effects in a direction very distant to the effect direction.

Higher power levels can be obtained if the weighting vector is chosen to be closer to the optimal. For this, the t^+ and t^+_{AD} permits the use of prior information and interim data. We next consider three scenarios with various prior information being available at the planning stage. Here, two-stage t^+_{AD} tests are considered with critical values set equal to $\alpha_{1,1} = 0.01$, $\alpha_{1,0} = 1$ and $\alpha_{2,1} = 0.0087$ ($\alpha = 0.05$).

The first scenario considers the case where no prior information is available and the investigators use the uniform weighting vector at the first stage. That is, the prior mean estimate is $\boldsymbol{m}_0 = \mathbf{1}_K$ and the prior covariance matrix estimate $\boldsymbol{S}_0$ has equal variances $s_0 = 1$ and correlations $r_0 = 0.7$, which result in angle of the first stage weighting vector, $\tilde{\boldsymbol{w}}_{t_1^+}$, to the optimal $\phi_{t_1^+} = 84.63° = \phi_{\mathbf{1}_K}$. The prior sample sizes are set $n_0 = \nu_0 = 1$, while the sample allocation $r_1 = 0.5$. For these values, the t^+ and t^+_{AD} result in a small increase to the power of the OLS, SS and PC tests, $\beta_{t^+} = 0.26$, $\beta_{t^+_{AD}} = 0.23$, explained by the decrease in the distance of $\tilde{\boldsymbol{w}}_{t_2^+}$, to the optimal (see figure 8.10) due to the interim adaptation. Despite the small increase, the latter power values are still very low implying a large chance of type II error.

The second scenario considers a more informative choice of prior estimates. This is perhaps more realistic than the first scenario, since most often in neuroimaging studies there is some prior information available about the direction of the effect. For instance, the knowledge of the ROI to be considered in the study can be useful in prior elicitation. There is substantial work in fMRI and other neuroimaging modalities about connectivity between ROI (see for example Alexander-Bloch et al. [2013]) and how to construct correlation matrices based on such results. Another way to construct correlation matrices is based on the location of ROI. The essence

of this method is that ROI which are at closer, in physical distance, locations are more highly correlated.

Here, we consider a prior covariance estimate constructed using the following simple model [Ripley, 1981; Worsley et al., 1991]

$$s_{0,kk'} = s_0^2 \exp(-cd_{kk'}), \qquad k,\ k' = 1, 2, \ldots, K, \tag{8.4}$$

with $s_0^2 = 1$ and $c = 0.002$ and $d_{kk'}$ the Euclidean distance between the centroids of $k-$ and $k'-$th ROI given in table 8.9. Note that for (bilateral) ROI with left and right areas, the ROI centroids with the smaller distance (nearest centroids) are used.

For prior covariance matrix as above and $\boldsymbol{m}_0 = \mathbf{1}_K$ ($n_0 = 0, \nu_0 = 4$), the angle of the first stage weighting vector to the optimal $\phi_{z_1^*} = 84.82°$ does not change substantially, compared to the uniform weighting vector, but the improvement of the direction of the angle due to the interim data is greater (see figure 8.10), which results in an increase of the power of the t^+ and t^+_{AD} tests, $\beta_{t^+} = 0.39$, $\beta_{t^+_{AD}} = 0.32$. Further power increase can be obtained, by using the available information about the connectivity between the above regions, but also the nature of the study which can reveal information about the mean structure. For instance, investigators are often able to predict which regions can potentially give larger effects based on the nature of the cognitive task (for example motor task, visual task) given to the subjects during the scan.

In our third hypothetical scenario, we consider the case where prior information suggests the same correlation matrix, but slightly lower activity in the ROI: AC, GP, T, VS. That is, $\boldsymbol{S}_0$ is same as in the second scenario ($\nu_0 = 4$) and the prior mean estimate $\boldsymbol{m}_0 = (0.5, 1, 1, 1, 0.5, 1, 1, 1, 0.5, 0.25, 1)^T$ ($n_0 = 7$). This decreases the angle $\phi_{t_1^+}$ and slightly the location of the distribution of the angles $\phi_{t_2^+}$ (see figure 8.10) resulting in a substantial power increase in the first stage of the study,

which gives $\beta_{t^+} = 0.42$ and $\beta_{t^+_{AD}} = 0.81$.

These hypothetical scenarios attempt to describe various settings that are likely to arise in fMRI studies. In the EEG study considered next, where more information about the study is available, a more realistic application of our tests is performed.

Table 8.9: Locations of the ROI centroids of the fMRI study.

ROI		x	y	z
Anterior Cingulate		0.715	19.183	24.387
Amygdala	Left	-23.039	-4.776	-18.399
	Right	23.328	-3.493	-18.399
Caudate	Left	-12.796	8.697	10.083
	Right	13.502	9.555	10.605
DLPFC		0.298	19.685	54.732
Globus Pallidus	Left	-19.139	-5.314	-1.331
	Right	19.854	-4.367	-1.346
Insular Cortex		0.921	2.258	-0.147
Orbitofrontal Cortex		-1.289	23.936	-16.046
Putamen	Left	-24.961	0.365	0.546
	Right	25.613	1.601	0.374
Substantia Nigra		9.000	-18.000	-13.000
Thalamus	Left	-10.200	-19.141	6.541
	Right	11.313	-18.126	6.870
Ventral Striatum	Left	-9.690	11.850	-6.952
	Right	9.597	12.678	-6.444

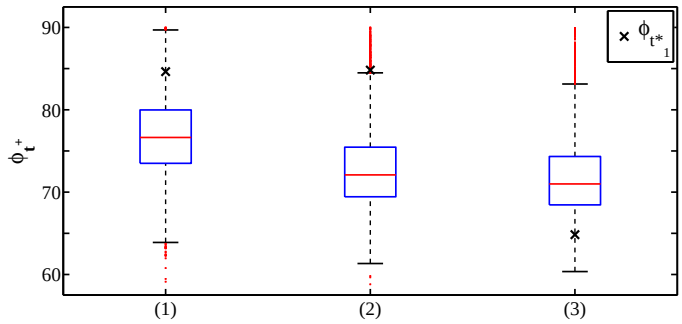


Figure 8.10: Box-plots of the distribution of the angles, $\phi_{t_2^+}$, of the second stage weighting vector, $\tilde{\boldsymbol{w}}_{t_2^+}$, to the optimal $\tilde{\boldsymbol{\omega}}^*$ at the three considered scenarios and the angle, $\phi_{t_1^+}$, of the first stage weighting vector, $\tilde{\boldsymbol{w}}_{t_1^+}$, to the optimal.

8.4.2 Application to the EEG study

As we have seen in section 2.4.2 (page 20), the EEG data provided in Läuter et al. [1996] represent the differences in the absolute theta power between the recordings at the beginning and at the end of a six week therapy given to $n_T = 19$ depressive patients. The $K = 9$ channels presented in figure 2.5 (page 20) were used.

Similarly to the previous section, we perform power analysis by setting the design parameters as in the above study, i.e. $K = 9$, $n_T = 19$, $\alpha = 0.05$ and the model parameters equal to their sample estimates $\boldsymbol{\mu} = \bar{\boldsymbol{x}}$, $\boldsymbol{\Sigma} = \boldsymbol{S}_x$ given in table 2.3 (page 21).

The Bonferroni global test, despite the increase in theta power of all channels, has low power $\beta_{BON} = 0.38$, suggesting small local effects. On the other hand, the relatively large value of the Mahalanobis distance, $\Delta = 1.15$, suggests strong global effects. Hotelling's T^2 test attains power $\beta_{T^2} = 0.68$, mainly limited by the sample size n_T being relatively close to the number of channels.

Taking into account that the single-stage t test for weighting vector equal to the optimal has power $\beta_{t^\star} = 0.9973$ we can easily understand that, similarly to the fMRI study, there is great scope for improvement, using linear combination tests. However, O'Brien's OLS and Läuter's SS and PC t tests take (very similar) low

power values, $\beta_{t_{OLS}} = 0.52$, $\beta_{t_{SS}} = 0.52$, $\beta_{t_{PC}} = 0.51$. Here, the uniform weighting vector of the OLS test, $\boldsymbol{w}_{OLS} \propto \mathbf{1}_K$, gives angle $\phi_{\boldsymbol{w}} \approx 71^\circ$ to the optimal.

Since the study was performed, there has been considerable research into EEG studies on depressive patients. There is now literature (see for example Davidson et al. [2002]; Stewart et al. [2011]) indicating that left-frontal hypoactivation and right-frontal hyperactivation are present in such subjects. This would indicate that a non-uniform prior over these frontal regions should be used. Using prior information based on such evidence, the adaptive t^+_{AD} test can attain high power levels.

For example, the prior estimates given in table 8.10 are in agreement with the evidence in the literature and further, the prior correlation structure is set to be roughly coherent to the distances between the channels, that is larger distances have smaller correlations, with larger correlations set at the highly active frontal regions (in accordance with the literature, see for example Lopes da Silva [2005]).

Table 8.10: Prior estimates of the means, standard deviations and correlations for the EEG depression study.

ch.(k)	3	4	5	6	7	8	17	18	19
$m_{0,k}$	0.5	3.5	1	2	2	2	2	2	2
$s_{0,k}$	1.5	2.5	1	2	2	2	2	2	2
$\boldsymbol{r}_{0,(3,k)}$	1								
$\boldsymbol{r}_{0,(4,k)}$	0.8	1							
$\boldsymbol{r}_{0,(5,k)}$	0.8	0.7	1						
$\boldsymbol{r}_{0,(6,k)}$	0.7	0.8	0.7	1					
$\boldsymbol{r}_{0,(7,k)}$	0.5	0.4	0.7	0.55	1				
$\boldsymbol{r}_{0,(8,k)}$	0.4	0.5	0.55	0.7	0.6	1			
$\boldsymbol{r}_{0,(17,k)}$	0.9	0.9	0.75	0.75	0.45	0.45	1		
$\boldsymbol{r}_{0,(18,k)}$	0.45	0.45	0.65	0.65	0.7	0.7	0.5	1	
$\boldsymbol{r}_{0,(19,k)}$	0.75	0.75	0.8	0.8	0.65	0.65	0.8	0.7	1

This prior estimate gives $\phi_{t_1^+} = 37.27°$ which is much smaller than the angle under the uniform weighting vector. For a two-stage design ($J = 2$), with balanced sample allocation, $n_1 = 10$, $n_2 = 9$, and α allocation $\alpha_{1,1} = 0.01$, $\alpha_{2,1} = 0.0087$, no early acceptance allowed, $\alpha_{0,1} = 1$, prior sample size $n_0 = 7 = 0.7n_1$, $\nu_0 = 6$ and the remaining design parameters as the original study, the t^+_{AD} test has power $\beta_{t^+_{AD}} = 0.84$ with $RSSR = 22.3\%$ ($E(n) = 15$). The t^+ test, for $n_0 = 6$, $n_1 = 6$, $n_2 = 13$ ($r = 0.3$) and the remaining design parameters as above, has power $\beta_{t^+} = 0.64$.

8.5 Discussion

In this chapter, we consider an extensive power analysis of various global tests, mainly focusing on the z^+, t^+ and adaptive z^+_{AD}, t^+_{AD} tests. In this we used the results of earlier chapters which allowed for relatively simple power analysis, despite the great complexities of multivariate tests.

A number of important outcomes are derived. The study of the relation of power to the sample and effect size confirms that z^+, t^+, z^+_{AD}, t^+_{AD} tests attain high power levels for any direction of the first stage weighting vector. Furthermore, our results indicate that balanced allocations of sample and α rate to the different stages of group sequential and adaptive designs are, in typical settings, more appropriate. The results of the impact of unknown $\mathbf{\Sigma}$ to the power of t^+_{AD} tests, reveals a few useful results about the behavior of this test, including that, for power analysis, it is often sufficient to study the special case of the first stage $\mathbf{\Sigma}-$deviation matrix being proportional to the identity matrix.

The comparison of various global tests show that the z^+, t^+ and adaptive z^+_{AD}, t^+_{AD}, in a similar way to the fully multivariate χ^2 and T^2 tests, can obtain high power levels for large enough sample or effect sizes, but they are particularly useful for small sample sizes, where the latter tests are inapplicable or inefficient. They lose in power compared to the linear combination z, t and z_{GS} and t_{GS} tests,

with fixed weighting vector, if the latter are close to the optimal, but they gain substantial power for distant to optimal weighting vectors.

Finally, the application to neuroimaging studies shows that the properties of this setting necessitate special consideration in global testing. The typically small and highly correlated local effects often make the Bonferroni global test inappropriate. Moreover, the small sample sizes do not permit the Hotelling's T^2 test to achieve high power levels. The developed linear combination tests can improve the power performance of these classical tests if prior knowledge and observed data set the targeted direction fairly close to the effect direction.

Chapter 9

Conclusions and future work

This thesis is devoted to the study of global testing procedures and particularly linear combination tests within the framework of single stage, group sequential and adaptive designs.

This work is motivated by the application of global testing in neuroimaging studies as introduced in chapter 2. Here, we discussed and illustrated through real examples, the demand for specialized statistical methodology taking into consideration some important characteristics typically observed in this field. First, neural measures of various brain locations are often highly correlated. Secondly, the treatment effects, if any, are commonly expressed in a dispersed fashion in the sense that they are locally small but, if combined, globally large. Thirdly, due to the great amount of literature in the field, the nature of the studies and the spatial characterization of the responses, prior information about the study is typically available. Finally, due to the high cost of the equipment and the need for trained individuals to run these studies, sample sizes are typically restricted to small levels. These sample sizes remain close to the observations' dimension even if the latter is reduced using methods such as ROI analysis for fMRI data and frequency analysis for EEG data.

Global hypotheses regarding the treatment effects across various brain locations are fundamental for neuroimaging studies. In chapter 3, we introduced the

main available approaches for global testing. We first discussed $p-$value adjustment methods, which are typically related to multiple testing methods such as the Bonferroni correction. We saw that these methods are typically simple and do not require strong modeling assumptions, but they target settings where one or some few independent local effects are present. In situations, such as our neuroimaging example studies, where correlations between the responses are relatively high and the effects are dispersed, these methods become inefficient and multivariate global tests are more appropriate. This is particularly the case when the multiple outcomes are biologically related, as in neuroimaging studies.

The fully multivariate χ^2 and Hotelling's T^2 tests were then discussed. We showed that these methods are efficient for sample sizes, n, substantially larger than the observation dimension, K, but the χ^2 test heavily relies on the knowledge of the covariance structure and Hotelling's T^2 becomes respectively inapplicable and inefficient if n is respectively smaller or larger but close to K. We underlined that this loss of efficiency, also shown in our application examples, is due to the search for effects in every direction of the multivariate space and we argue that this loss can be avoided if prior information about the direction of the effect is available. This can be practically implemented by linear combination tests where the targeted direction is selected using the weighting vector. The latter reduces the observation vectors to scalar linear combinations which are then used to construct z and t test statistics. The available approaches for linear combination testing were discussed, including their shortcoming of requiring specific effect structures, such as the uniform, to attain high power performance.

Taking into consideration the properties of neuroimaging studies and that the available global testing methodology is often not able to accommodate them, in chapter 4, we developed the predictive-power-optimal linear combination z^+ and t^+ tests. We first derived the power-optimal z^* and t^* tests maximizing the power of linear combination z and t tests (for fixed values of the modeling parameters) with

respect to the weighting vector $\boldsymbol{w}$. This allowed us to derive some results related to existing methodology, including the cases where O'Brien's tests are power-optimal and a comparison of t^* with Hotelling's T^2 test. In the latter, we proved that the power of the t^* is always larger than the power of T^2 test. A comparison of the power of the two tests, reveals the great potential of linear combination tests to improve power performance over T^2, especially when n is close to K.

As the z^* and t^* tests are performed using the optimal but unknown weighting vector $\boldsymbol{\omega}^*$, methodology to obtain the weighting vector using prior information and pilot data was proposed. The resulting z^+ and t^+ tests are optimal, within the class of linear combination tests, with respect to predictive power given the available information. The z^+ and t^+ tests, which are analytically proved to control type I error, satisfy our requirements for application to neuroimaging studies. Firstly, they incorporate correlations into the multivariate normality assumption which allows us to capture globally large but dispersed effects. Secondly, they provide a framework for formally incorporating prior information about the effect structure in selecting the targeted direction and thirdly, as it becomes more clear in power analysis, they are appropriate for small sample sizes.

The use of preliminary data obtained from an early pilot study allows for potentially correcting imprecise directions of the weighting vector, particularly if the latter is nearly orthogonal to the effect direction. In chapter 5, we introduced the framework of adaptive designs which allows us to replace this external pilot study, with an internal pilot study which is then considered as the first stage of a two-stage or, generally, multi-stage adaptive testing procedure. We discussed how the concepts of monitoring data acquisition with the possibility for early termination and interim design modifications give rise to sequential and adaptive designs and we introduced methodology for performing group sequential and adaptive testing. Chapter 5 closes with a discussion about the potential and challenges to be overcome from these designs and particularly adaptive designs. We argued that great effort

for careful planning should be made in adaptive designs to acquire relevant prior information to be used in planning the study. We also disputed that, not only flexibility and type I error control, but also power performance should be targeted and carefully studied when using adaptive designs. Optimal pre-specified adaptation rules can be important towards this direction.

In chapter 6, the methodology of adaptive designs was used to perform global testing. The adaptive $J-$stage ($J \geq 2$) linear combination z^{+}_{AD} and t^{+}_{AD} tests sequentially adapt the weighting vectors, initially selected using prior information, based on the information available at each interim analysis. The adaptation rule is optimal in terms of the predictive power given the results available at each interim point.

The developed tests, which are analytically proved to control type I error, have similar properties to z^{+} and t^{+}, including those related to the application to neuroimaging studies, but they allow some important additional possibilities. Firstly, they can be implemented as a multi-stage test with the option of early stopping in settings where such procedures are considered as more appropriate. Secondly, the first stage data is not only used for selecting the weighting vector, which in settings with small sample sizes might be considered as inappropriate, but also as observations used for testing. This can also be potentially beneficial in terms of power performance especially in cases where the prior estimates set the targeted direction relatively close to the effect direction.

In chapter 7, a characterisation of the power function of linear combination tests was derived. This was useful to understand the behaviour of linear combination tests and explore their power performance by reducing their potentially high-dimensional design space to a lower dimensional space described by easily interpretable summary measures. In particular, these results showed that the power of linear combination tests is determined by two measures. The first is the strength of the treatment effect described by the well-known scalar Mahalanobis distance.

The second is the distance between the effect direction and the targeted direction, determined by the weighting vector. For the single-stage z, t and group sequential z_{GS} and t_{GS} linear combination tests with fixed weighting vectors as well as the z^+ and adaptive z^+_{AD} this distance is measured by a scalar measure, the angle of the selected weighting vector to the optimal. For the t^+ and t^+_{AD} tests, this distance is more complex, but methods have been developed which improve our understanding to the behavior of these tests and reduce substantially the dimension to be considered in their power analysis.

The results of this power characterisation are used to perform power analysis of linear combination tests, the main results of which were presented in chapter 8. Here we studied the power performance of z^+, t^+ and the adaptive z^+_{AD} and t^+_{AD} tests as well as the single-stage z, t and χ^2, T^2 tests and their group sequential counterparts z_{GS}, t_{GS} and χ^2_{GS}, T^2_{GS} tests. These empirical studies showed that the above tests attain power 1 for large sample sizes and/or large effect sizes, with exception to the z, t, z_{GS} and t_{GS} tests with weighting vector fixed orthogonal to the optimal in which case power is equal to the significance level, α. Furthermore, the results implied that balanced allocations of the sample size and the α rate to the different stages of multi-stage tests are, in most cases, more appropriate regarding power and average sample size.

In terms of comparisons, the z, t and z_{GS}, t_{GS} gain considerable amount of power compared to z^+, t^+, z^+_{AD}, t^+_{AD} tests for weighting vectors very close to the optimal. However, for such small angles the latter tests also achieve good power performance, while the interim re-assessment of the weighting vector in the latter tests can substantially reduce large angles of the initial weighting vector to the optimal. High power levels can therefore be potentially attained with z^+, t^+, z^+_{AD} and t^+_{AD} even if the initial weighting vector is orthogonal to the optimal. The fully multivariate χ^2 and χ^2_{GS} tests attain relatively high power levels (in our simulations close to the power of z^+, z^+_{AD} with initial weighting vector around 45° from optimal).

Similar levels are also attained by the T^2 and T^2_{GS} tests for large sample sizes, but for sample sizes close to the observation dimension their power performance is greatly affected, taking values close to the t^+ and t^+_{AD} tests with nearly orthogonal angles of the initial weighting vector to the optimal.

The application to the real examples of the fMRI drug development study and the EEG depression study illustrated situations where the existing methodology for global tests cannot provide acceptable power performance for various different reasons. It also showed that the developed linear combination tests can provide good power performance under some conditions. For the small sample sizes presented in these examples, the selection of the weighting vector is greatly affected by the prior estimates. However, our application shows that, if prior information is available (a situation which is not rare in neuroimaging studies), sufficient power levels can be attained with t^+ and especially with the adaptive t^+_{AD} test.

While the work presented in the thesis satisfies some of the targets of global testing, especially within the motivational neuroimaging setting, plenty of questions, generalisations and extensions remain to be explored. In the following, some of these areas of future research are discussed.

9.1 Future work

The methodology developed in this thesis, but also, generally, the best part of existing global testing methodology, is based on the assumption of multivariate normality. As we discussed in section 3.3 (see page 30), there are various reasons why this assumption is central in multivariate statistics (for example, the Central Limit Theorem). Furthermore, normality is the dominant assumption in neuroimaging and specifically fMRI and EEG. However, multivariate normality in real applications is always an approximation at best and if substantial departures from normality are foreseen, a different set of assumptions might be more appropriate.

Particularly, if sharper peaks than the normal distribution are foreseen, a distribution such as the double exponential might be preferred [Mangoubi et al., 2008]. In this case, despite the extra complexities to be encountered in calculating the test properties, the methodology for deriving the linear combination test statistic and particularly the weighting vector, is expected to provide similar results. On the contrary, in situations where multi-modal or skewed distributions are anticipated [Azzalini and DallaValle, 1996; Ferreira and Steel, 2007], a single weighting vector is expected to not be sufficient to capture the data variability and describe the effect direction.

In the latter situation, a generalisation of the present methodology where, not necessarily a single, but a number of weighting vectors are selected at the interim analyses of the $J-$stage adaptive test, might be necessary. Here, an optimization, in terms of the direction, but also the number of selected weighting vectors, taking into account data variability and generally the form of the observations' distribution, needs to be considered. If the optimization method is effective, these more flexible procedures should deliver higher efficiency and robustness.

Furthermore, a substantial generalization of the proposed methodology might be achieved by extending the search for optimal tests beyond the class of linear combination tests to the wider class of multivariate tests. Bayesian methods and decision theory can be useful to this end.

The developed methodology has links with hypothesis testing problems occurring in multivariate change-point analysis. Multivariate change-point analysis targets on detecting location shifts of $K-$dimensional time-series by testing against stationarity [Chen and Gupta, 2012]. Similarly to the cross-sectional setting considered in this thesis, it is often preferable or essential, especially for large K, to perform these tests by first reducing the multivariate time-series to low-dimensional summary time-series. An extension of the present methodology may consider new methods for performing multivariate change-point analysis based on optimal dimen-

sion reduction operators. In settings, as in neuroimaging, where prior information is typically available, these dimension reduction operators can be derived using this information and accumulated data. These methods can potentially improve substantially the performance of currently available tests such as the one proposed as part of the methodology developed in Aston and Kirch [2012].

The methods developed so far are mainly suitable for observations of moderate dimension. In many important statistical applications, including genetics and neuroimaging, the dimension of observations, if not reduced as a preparatory step, is extremely large. Therefore, it is often preferable to treat the observations as functional. Functional data analysis is particularly attractive in neuroimaging, since it allows the neural activity of the whole brain to be naturally represented as a single function [Aston and Kirch, 2012]. An important target of functional data analysis is testing hypotheses regarding the mean of functional observations [Ramsay and Silverman, 2005]. In this setting, it is often essential to construct test statistics using low dimensional summaries of the functional observations. An extension of the methodology presented in this thesis may explore new methods for performing dimensionality reductions and hypothesis testing in functional data analysis.

Appendix A

Proofs of lemma 6.3.1 and 6.3.2 for t tests

We next prove that analogous to the results, regarding the z test, described in lemmas 6.3.1, 6.3.2 (pp. 90, 93) hold also for the t test. This leads to the proof of lemma 6.3.3 (page 94) regarding the t test. The methods of the proofs are very similar to above, but for clarification and to avoid confusion due to the complexity introduced by the non-centrally t distributed test statistics, the proofs are provided in complete form.

Proof of lemma 6.3.1 *for the* t *test.* For convenience we use in this proof the notation $\tilde{t} = xt/\sqrt{\nu_1}$. For the first stage power we have that

$$\frac{\mathrm{d}\beta_{t_1}}{\mathrm{d}\theta_1} = \int_0^\infty \sqrt{n_1} \left(\phi \left(\tilde{t}_{\nu_1, \alpha_{1,1}/2} - \bar{\theta}_1 \right) - \phi \left(\tilde{t}_{\nu_1, \alpha_{1,1}/2} + \bar{\theta}_1 \right) \right) h_{\nu_1}(x) \mathrm{d}x. \tag{A.1}$$

The second stage power can be written as

$$\beta_{t_2} = \int_{\mathcal{C}_{t_1}^+} \beta_{c,t_2}(\theta_2; t_1) \left(\psi_{\bar{\theta}_1, \nu_1}(t_1) + \psi_{\bar{\theta}_1, \nu_1}(-t_1) \right) \mathrm{d}t_1,$$

where $\mathcal{C}_{t_1}^{+} = \left(t_{\nu_1,\alpha_{0,1}/2}, t_{\nu_1,\alpha_{1,1}/2}\right)$. Thus,

$$\frac{\mathrm{d}\beta_{t_2}}{\mathrm{d}\theta_1} = \int_{\mathcal{C}_{t_1}^{+}} \beta_{c,t_2}(\theta_2; t_1) \frac{\mathrm{d}\left(\psi_{\bar{\theta}_1,\nu_1}(t_1) + \psi_{\bar{\theta}_1,\nu_1}(-t_1)\right)}{\mathrm{d}\theta_1} \mathrm{d}t_1. \tag{A.2}$$

Using (4.10), we can re-write the derivative,

$$\begin{aligned}
&\frac{\mathrm{d}\left(\psi_{\bar{\theta}_1,\nu_1}(t_1) + \psi_{\bar{\theta}_1,\nu_1}(-t_1)\right)}{\mathrm{d}\theta_1} \\
=&\int_0^{\infty} \sqrt{n_1} \left(\left(\tilde{t}_1 - \bar{\theta}_1\right) \phi\left(\tilde{t}_1 - \bar{\theta}_1\right) - \left(\tilde{t}_1 + \bar{\theta}_1\right) \phi\left(\tilde{t}_1 + \bar{\theta}_1\right)\right) h_{\nu_1}(x) \mathrm{d}x \\
=&\int_0^{\infty} \frac{x\sqrt{n_1}}{\sqrt{\nu_1}} \left(e^{2\bar{\theta}_1\tilde{t}_1} + 1\right) \phi\left(\tilde{t}_1 + \bar{\theta}_1\right) \left(\tilde{t}_1 \tanh\left(\bar{\theta}_1\tilde{t}_1\right) - \bar{\theta}_1\right) h_{\nu_1}(x) \mathrm{d}x.
\end{aligned} \tag{A.3}$$

First note that

$$\frac{\mathrm{d}\tanh\left(\bar{\theta}_1\tilde{t}_1\right)}{\mathrm{d}t_1} = \frac{4\bar{\theta}_1\tilde{t}_1 e^{2\bar{\theta}_1\tilde{t}_1}}{\left(e^{2\bar{\theta}_1\tilde{t}_1} + 1\right)^2}, \tag{A.4}$$

which is greater or equal to zero for $\theta_1 \geq 0$ and negative for $\theta_1 < 0$. Using the last result, we can easily see that

$$\frac{\mathrm{d}}{\mathrm{d}t_1}\left(\tilde{t}_1 \tanh\left(\bar{\theta}_1\tilde{t}_1\right) - \bar{\theta}_1\right)$$

is non-negative for $\theta_1 \geq 0$ and negative for $\theta_1 < 0$ ($t_1 > 0$). Thus, for $\theta_1 \geq 0$, we can identify the following three cases:

$$\begin{aligned}
&(1) \qquad \tilde{t}_1 \tanh\left(\bar{\theta}_1\tilde{t}_1\right) - \bar{\theta}_1 \geq 0, \ \forall\, t_1 \in \mathcal{C}_{t_1}^{+}, \\
&(2) \qquad \tilde{t}_1 \tanh\left(\bar{\theta}_1\tilde{t}_1\right) - \bar{\theta}_1 \leq 0, \ \forall\, t_1 \in \mathcal{C}_{t_1}^{+}, \\
&(3) \qquad \exists\, \dot{t}_1 \in \mathcal{C}_{t_1}^{+} : \tilde{t}_1 \tanh\left(\bar{\theta}_1\tilde{t}_1\right) - \bar{\theta}_1 \leq 0, \ \forall\, t_1 \in \left(t_{\nu_1,\frac{\alpha_{0,1}}{2}}, \dot{t}_1\right), \\
&\qquad\qquad\qquad\qquad \tilde{t}_1 \tanh\left(\bar{\theta}_1\tilde{t}_1\right) - \bar{\theta}_1 > 0, \ \forall\, t_1 \in \left(\dot{t}_1, t_{\nu_1,\frac{\alpha_{1,1}}{2}}\right).
\end{aligned}$$

In case (1), by (A.3), we have that

$$\frac{\mathrm{d}(\psi_{\bar{\theta}_1,\nu_1}(t_1) + \psi_{\bar{\theta}_1,\nu_1}(-t_1))}{\mathrm{d}\theta_1} \geq 0, \quad \forall\, t_1 \in \mathcal{C}_{t_1}^{+}.$$

Therefore, by (A.2), β_{t_2} and hence β_t is non-negative for $\theta_1 \geq 0$.

In case (2), by (A.3), we have that

$$\frac{\mathrm{d}(\psi_{\bar{\theta}_1,\nu_1}(t_1) + \psi_{\bar{\theta}_1,\nu_1}(-t_1))}{\mathrm{d}\theta_1} \leq 0, \ \ \forall\, t_1 \in \mathcal{C}^+_{t_1}.$$

The last result using (A.2) in turn implies that

$$\frac{\mathrm{d}\beta_{t_2}}{\mathrm{d}\theta_1} \geq \beta_{c,t_2}(\theta_2; t_{\nu_1,\frac{\alpha_{1,1}}{2}}) \int_{t_{\nu_1,\frac{\alpha_{0,1}}{2}}}^{t_{\nu_1,\frac{\alpha_{1,1}}{2}}} \frac{\mathrm{d}\left(\psi_{\bar{\theta}_1,\nu_1}(t_1) + \psi_{\bar{\theta}_1,\nu_1}(-t_1)\right)}{\mathrm{d}\theta_1} \mathrm{d}t_1 \geq 0.$$

Therefore, using (A.1) we derive that

$$\begin{aligned}\frac{\mathrm{d}\beta_t}{\mathrm{d}\theta_1} \geq \left(1 - \beta_{c,t_2}\left(\theta_2; t_{\nu_1,\frac{\alpha_{1,1}}{2}}\right)\right) \frac{\mathrm{d}}{\mathrm{d}\theta_1}\left(\Psi_{\bar{\theta}_1,\nu_1}(-t_{\nu_1,\frac{\alpha_{1,1}}{2}}) - \Psi_{\bar{\theta}_1,\nu_1}(t_{\nu_1,\frac{\alpha_{1,1}}{2}})\right) \\ + \beta_{c,t_2}\left(\theta_2; t_{\nu_1,\frac{\alpha_{1,1}}{2}}\right) \frac{\mathrm{d}}{\mathrm{d}\theta_1}\left(\Psi_{\bar{\theta}_1,\nu_1}(-t_{\nu_1,\frac{\alpha_{0,1}}{2}}) - \Psi_{\bar{\theta}_1,\nu_1}(t_{\nu_1,\frac{\alpha_{0,1}}{2}}))\right) \geq 0.\end{aligned}$$

In case (3), by (A.3), we have that

$$\begin{aligned}\frac{\mathrm{d}(\psi_{\bar{\theta}_1,\nu_1}(t_1) + \psi_{\bar{\theta}_1,\nu_1}(-t_1))}{\mathrm{d}\theta_1} \leq 0, &\qquad \forall\, t_1 \in (t_{\nu_1,\frac{\alpha_{0,1}}{2}}, \dot{t}_1) \\ \frac{\mathrm{d}(\psi_{\bar{\theta}_1,\nu_1}(t_1) + \psi_{\bar{\theta}_1,\nu_1}(-t_1))}{\mathrm{d}\theta_1} > 0, &\qquad \forall\, t_1 \in (\dot{t}_1, t_{\nu_1,\frac{\alpha_{1,1}}{2}}).\end{aligned}$$

Thus, using (A.2), we derive that

$$\begin{aligned}\frac{\mathrm{d}\beta_{t_2}(\theta_1)}{\mathrm{d}\theta_1} \geq \beta_{c,t_2}(\theta_2; \dot{t}_1) \int_{t_{\nu_1,\frac{\alpha_{0,1}}{2}}}^{\dot{t}_1} \frac{\mathrm{d}(\psi_{\nu_1}(t_1;\theta_1) + \psi_{\nu_1}(-t_1;\theta_1))}{\mathrm{d}\theta_1} \mathrm{d}t_1 \\ + \beta_{c,t_2}(\theta_2; \dot{t}_1) \int_{\dot{t}_1}^{t_{\nu_1,\frac{\alpha_{1,1}}{2}}} \frac{\mathrm{d}(\psi_{\nu_1}(t_1;\theta_1) + \psi_{\nu_1}(-t_1;\theta_1))}{\mathrm{d}\theta_1} \mathrm{d}t_1 \geq 0.\end{aligned}$$

Hence, using (A.1), we have that

$$\begin{aligned}\frac{\mathrm{d}\beta_t(\theta_1)}{\mathrm{d}\theta_1} \geq \left(1 - \beta_{c,t_2}(\theta_2; \dot{t}_1)\right) \frac{d}{d\theta_1}\left(\Psi_{\nu_1}(-t_{\nu_1,\frac{\alpha_{1,1}}{2}};\theta_1) - \Psi_{\nu_1}(t_{\nu_1,\frac{\alpha_{1,1}}{2}};\theta_1)\right) \\ + \beta_{c,t_2}\left(\theta_2; \dot{t}_1\right) \frac{d}{d\theta_1}\left(\Psi_{\nu_1}(-t_{\nu_1,\frac{\alpha_{0,1}}{2}};\theta_1) - \Psi_{\nu_1}(t_{\nu_1,\frac{\alpha_{0,1}}{2}};\theta_1))\right) \geq 0.\end{aligned}$$

In all three possible cases the power, β_t, of the two-stage t test is non-decreasing in θ_1 for $\theta_1 \geq 0$. Similarly, we can prove that β_t is decreasing in θ_1 for $\theta_1 < 0$ and the result follows. □

Proof of lemma 6.3.2 *for the t test.* The second stage power, β_{t_2}, can be written as

$$\beta_{t_2} = \int_{\mathcal{C}_{t_1}} \beta_{c,t_2}(\theta_2; t_1)\psi_{\bar{\theta}_1,\nu_1}(t_1)\mathrm{d}t_1,$$

where $\mathcal{C}_{t_1} = \left(-t_{\nu_1,\alpha_{1,1}/2}, -t_{\nu_1,\alpha_{0,1}/2}\right) \cup \left(t_{\nu_1,\alpha_{0,1}/2}, t_{\nu_1,\alpha_{1,1}/2}\right)$ and the second stage conditional power

$$\beta_{c,t_2}(\theta_2; t_1) = Pr\left(|t_2| \geq t_{\nu_2,\frac{\alpha'_{1,2}}{2}} \mid p_{t_1}\right) = 1 - \Psi_{\bar{\theta}_2,\nu_2}\left(t_{\nu_2,\frac{\alpha'_{1,2}}{2}}\right) + \Psi_{\bar{\theta}_2,\nu_2}\left(-t_{\nu_2,\frac{\alpha'_{1,2}}{2}}\right).$$

Using equation (4.10) we derive that

$$\frac{\mathrm{d}\beta_{c,t_2}(\theta_2; t_1)}{\mathrm{d}\theta_2} = \int_0^\infty \sqrt{n_2}\left(\phi\left(\frac{xt_{\nu_2,\alpha'_{1,2}/2}}{\sqrt{\nu_2}} - \bar{\theta}_2\right) - \phi\left(\frac{xt_{\nu_2,\alpha'_{1,2}/2}}{\sqrt{\nu_2}} + \bar{\theta}_2\right)\right) h_{\nu_2}(x)\mathrm{d}x,$$

which is non-negative for $\theta_2 \geq 0$ and negative for $\theta_2 < 0$. Therefore, the conditional power β_{c,t_2} and hence the power of the two-stage t test is non-decreasing in $|\theta_2|$. □

Bibliography

Adolf, D., Baecke, S., Kahle, W., Bernarding, J., and Kropf, S. (2011). Applying multivariate techniques to high-dimensional temporally correlated fMRI data. *Journal of Statistical Planning and Inference*, 141(12):3760–3770.

Alexander-Bloch, A., Giedd, J. N., and Bullmore, E. (2013). Imaging structural covariance between human brain regions. *Nature Reviews Neuroscience*, 14(5):322–336.

Almeida, R. and Ledberg, A. (2002). Exact multivariate tests for brain imaging data. *Human Brain Mapping*, 16(1):24–35.

Anderson, K. (2013). *gsDesign, Version 2.7-06 [computer software]*. Merck Research Laboratories.

Anderson, T. W. (1984). *An introduction to multivariate statistical analysis*. Wiley, New York.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*. Wiley-Interscience, Hoboken, N.J.

Aquilina, K., Edwards, P., and Strong, A. (2005). Principles and practice of image-guided neurosurgery. In Moore, A. and Newell, D., editors, *Neurosurgery: Principles and Practice*, pages 123–138–. Springer London.

Armitage, P. (1954). Sequential tests in prophylactic and therapeutic trials. *The Quarterly Journal of Medicine*, 23(91):255–74.

Armitage, P. (1958). Sequential-methods in clinical-trials. *American Journal of Public Health and the Nations Health*, 48(10):1395–1402.

Armitage, P., Berry, G., and Matthews, J. N. S. (2002). *Statistical methods in medical research*. Blackwell Science.

Aston, J. A. D. and Kirch, C. (2012). Evaluating stationarity via change-point alternatives with applications to fMRI data. *Annals of Applied Statistics*, 6(4):1906–1948.

Azzalini, A. and DallaValle, A. (1996). The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726.

Barber, S. and Jennison, C. (2002). Optimal asymmetric one-sided group sequential tests. *Biometrika*, 89(1):49–60.

Bauer, G. and Bauer, R. (2005). EEG, drug effects, and central nervous system poisoning. In Niedermeyer, E. and Lopes da Silva, F. H., editors, *Electroencephalography, 5th edition*. Lippincott Williams and Wilkins, Philadelphia.

Bauer, P. and Einfalt, J. (2006). Application of adaptive designs - a review. *Biometrical Journal*, 48(4):493–506.

Bauer, P. and Köhne, K. (1994). Evaluation of experiments with adaptive interim analyses. *Biometrics*, 50(4):1029–1041.

Bear, M. F., Connors, B. W., and Paradiso, M. A. (2007). *Neuroscience : Exploring the brain*. Lippincott Williams & Wilkins, Philadelphia, PA.

Birkett, M. A. and Day, S. J. (1994). Internal pilot studies for estimating sample size. *Statistics in Medicine*, 13(23-24):2455–2463.

Bittman, R. M., Romano, J. P., Vallarino, C., and Wolf, M. (2009). Optimal testing of multiple hypotheses with common effect direction. *Biometrika*, 96(2):399–410.

Borsook, D., Hargreaves, R. J., and Becerra, L. (2011). Can functional magnetic resonance imaging improve success rates in central nervous system drug discovery? *Expert Opinion on Drug Discovery*, 6(6):597–617.

Brannath, W., Gutjahr, G., and Bauer, P. (2012). Probabilistic foundation of confirmatory adaptive designs. *Journal of the American Statistical Association*, 107(498):824–832.

Brannath, W., Koenig, F., and Bauer, P. (2007). Multiplicity and flexibility in clinical trials. *Pharmaceutical Statistics*, 6(3):205–216.

Brannath, W., Posch, M., and Bauer, P. (2002). Recursive combination tests. *Journal of the American Statistical Association*, 97(457):236–244.

Brannath, W., Zuber, E., Branson, M., Bretz, F., Gallo, P., Posch, M., and Racine-Poon, A. (2009). Confirmatory adaptive designs with Bayesian decision tools for a targeted therapy in oncology. *Statistics in Medicine*, 28(10):1445–1463.

Bretz, F., Koenig, F., Brannath, W., Glimm, E., and Posch, M. (2009). Adaptive designs for confirmatory clinical trials. *Statistics in Medicine*, 28(8):1181–1217.

Bretz, F., Schmidli, H., König, F., Racine, A., and Maurer, W. (2006). Confirmatory seamless phase II/III clinical trials with hypotheses selection at interim: General concepts. *Biometrical Journal*, 48(4):623–634.

Burman, C.-F. and Sonesson, C. (2006). Are flexible designs sound? *Biometrics*, 62(3):664–669.

Burneo, J., Demaerschalk, B., and Jenkins, M. (2011). *Neurology: An evidence-based approach*. Springer.

Cabeza, R., Kingstone, A., and Bub, D. (2001). Handbook of functional neuroimaging of cognition. *Canadian Psychology*, 42(4):318–320.

Chang, M. (2007). Adaptive design method based on sum of p-values. *Statistics in Medicine*, 26(14):2772–2784.

Chen, J. and Gupta, A. K. (2012). *Parametric statistical change point analysis with applications to genetics, medicine, and finance*. Birkhäuser, Boston.

Chi, L., Hung, H. M. J., and Wang, S.-J. (1999). Modification of sample size in group sequential clinical trials. *Biometrics*, 55(3):853–857.

Chuong-Stein, C., Anderson, K., Gollo, P., and Collins, S. (2006). Sample size reestimation : A review and recommendations. *Drug Information Association*, 40(4):10.

Collura, T. F., Thatcher, R. W., Smith, M. L., Lambos, W. A., and Stark, C. R. (2009). Chapter 5 - EEG biofeedback training using live Z-scores and a normative database. In Budzynski, T. H., Budzynski, H. K., Evans, J. R., and Abarbanel, A., editors, *Introduction to Quantitative EEG and Neurofeedback (Second Edition)*, pages 103–141. Academic Press, San Diego.

Cui, L. and Wu, X. (2010). Sample size reestimation design with applications in clinical trials. In *Handbook of Adaptive Designs in Pharmaceutical and Clinical Development*, pages 7–1–7–16–. CRC Press.

D'Agostino, R. B. and Russell, H. K. (2005). *Multiple endpoints, multivariate global tests*. John Wiley & Sons, Ltd.

Dallow, N. S., Leonov, S. L., and Roger, J. H. (2008). Practical usage of O'Brien's OLS and GLS statistics in clinical trials. *Pharmaceutical Statistics*, 7(1):53–68.

Davidson, R. J., Pizzagalli, D., Nitschke, J. B., and Putnam, K. (2002). Depression: Perspectives from affective neuroscience. *Annual Review of Psychology*, 53:545–574.

Denne, J. S. and Jennison, C. (2000). A group sequential t-test with updating of sample size. *Biometrika*, 87(1):125–134.

Dmitrienko, A., Tamhane, A. C., and Bretz, F. (2010). *Multiple testing problems in pharmaceutical statistics*. Chapman & Hall/CRC, Boca Raton, FL.

Eales, J. D. (1995). Optimal two-sided group sequential tests. *Sequential Analysis*, 14(4):273–286.

Eales, J. D. and Jennison, C. (1992). An improved method for deriving optimal one-sided group sequential tests. *Biometrika*, 79(1):13–24.

EMEA (2007). Reflection paper on methodological issues in confirmatory clinical trials with flexible design and analysis plan, committee for medicinal products for human use (CHMP), European Medicine Agency.

Fang, K.-t. and Zhang, Y.-T. (1990). *Generalized multivariate analysis*. Science Press ; Springer-Verlag, Beijing; Berlin; New York.

FDA (2010). Guidance for industry-adaptive design clinical trials for drugs and biologics, The U.S. Food and Drug Administration.

Ferreira, J. T. A. S. and Steel, M. F. J. (2007). A new class of skewed multivariate distributions with applications to regression analysis. *Statistica Sinica*, 17(2):505–529.

Fisher, R. A. (1970). *Statistical methods for research workers.* Oliver and Boyd, Edinburgh.

Follmann, D. (1996). A simple multivariate test for one-sided alternatives. *Journal of the American Statistical Association*, 91(434):854–861.

Frick, H. (1997). A note on the bias of O'Brien's OLS test. *Biometrical Journal*, 39(1):125–128.

Friedman, L. M., Furberg, C., and DeMets, D. L. (2010). *Fundamentals of clinical trials*. Springer, New York.

Friston, K., Ashburner, J., Kiebel, S., Nichols, T., and Penny, W. (2007). *Statistical parametric mapping : The analysis of funtional brain images*. Elsevier/Academic Press, Amsterdam; Boston.

Friston, K. J. (2009). Modalities, modes, and models in functional neuroimaging. *Science*, 326(5951):399–403.

Gallo, P., Chuang-Stein, C., Dragalin, V., Gaydos, B., Krams, M., and Pinheiro, J. (2006). Adaptive designs in clinical drug developmentan executive summary of the PhRMA working group. *Journal of Biopharmaceutical Statistics*, 16(3):275–283.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). *Bayesian data analysis*. Chapman & Hall/CRC, Boca Raton, Fla.

George, E. O. and Mudholkar, G. S. (1990). P-values for two-sided tests. *Biometrical Journal*, 32(6):747–751.

Gevins, A. (2002). 8 - Electrophysiological imaging of brain function. In Toga, A. W. and Mazziotta, J. C., editors, *Brain Mapping: The Methods (Second Edition)*, pages 175–188. Academic Press, San Diego.

Gould, A. L. (2001). Sample size re-estimation: Recent developments and practical considerations. *Statistics in Medicine*, 20(17-18):2625–2643.

Gupta, S. D. and Perlman, M. D. (1974). Power of the noncentral F-test: Effect of additional variates on Hotelling's T^2-test. *Journal of the American Statistical Association*, 69(345):174–180.

Handy, T. C. (2005). *Event-related potentials : A methods handbook*. MIT Press, Camridge, Mass.

Hauk, O. (2013). Introduction to EEG and MEG. *http://imaging.mrc-cbu.cam.ac.uk/meg/IntroEEGMEG*.

Hedges, L. V. and Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press, Orlando.

Hemmelmann, C., Horn, M., Reiterer, S., Schack, B., Susse, T., and Weiss, S. (2004). Multivariate tests for the evaluation of high-dimensional EEG data. *Journal of Neuroscience Methods*, 139(1):111–120.

Hommel, G. (1988). A stagewise rejective multiple test procedure based on a modified Bonferroni test. *Biometrika*, 75(2):383–386.

Hommel, G. (2001). Adaptive modifications of hypotheses after an interim analysis. *Biometrical Journal*, 43(5):581–589.

Honey, G. and Bullmore, E. (2004). Human pharmacological MRI. *Trends in Pharmacological Sciences*, 25(7):366 – 374.

Hotelling, H. (1931). The generalization of student's ratio. *The Annals of Mathematical Statistics*, 2(3):360–378.

Hu, F. and Rosenberger, W. F. (2003). Optimality, variability, power: Evaluating response-adaptive randomization procedures for treatment comparisons. *Journal of the American Statistical Association*, 98(463):671–678.

Hu, F. F. and Zhang, L. X. (2004). Asymptotic properties of doubly adaptive biased coin designs for multitreatment clinical trials. *Annals of Statistics*, 32(1):268–301.

Huettel, S. A., Song, A. W., and McCarthy, G. (2008). *Functional magnetic resonance imaging*. Sinauer Associates, Sunderland, Mass.

Huson, L. (2009). The Bayesian bootstrap in a predictive power analysis. *Case Studies in Business, Industry and Government Statistics*, 3:18–22.

Hwang, I. K., Shih, W. J., and De Cani, J. S. (1990). Group sequential designs using a family of type I error probability spending functions. *Statistics in Medicine*, 9(12):1439–1445.

Iannetti, G. and Wise, R. G. (2007). BOLD functional MRI in disease and pharmacological studies: Room for improvement? *Magnetic Resonance Imaging*, 25(6):978 – 988.

ICH (1998). International Harmonised Tripartite Guideline: Statistical principles for clinical trials E9. http://www.ich.org/LOB/media/MEDIA485.pdf.

James, S. (1991). Approximate multinormal probabilities applied to correlated multiple end-points in clinical-trials. *Statistics in Medicine*, 10(7):1123–1135.

Jennison, C. and Turnbull, B. W. (1990). Statistical approaches to interim monitoring of medical trials: A review and commentary. *Statistical Science*, 5(3):299–317.

Jennison, C. and Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Chapman & Hall/CRC, Boca Raton.

Jennison, C. and Turnbull, B. W. (2001). Group sequential tests with outcome-dependent treatment assignment. *Sequential Analysis*, 20(4):209–234.

Jennison, C. and Turnbull, B. W. (2003). Mid-course sample size modification in clinical trials based on the observed treatment effect. *Statistics in Medicine*, 22(6):971–993.

Jennison, C. and Turnbull, B. W. (2006). Efficient group sequential designs when there are several effect sizes under consideration. *Statistics in Medicine*, 25(6):917–932.

Jennison, C. and Turnbull, B. W. (2007). Adaptive seamless designs: Selection and prospective testing of hypotheses. *Journal of Biopharmaceutical Statistics*, 17(6):1135–1161.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous univariate distributions. Norman L. Johnson, Samuel Kotz, N. Balakrishnan. Vol. 2*. Wiley & Sons, New York; Chichester.

Kieser, M. and Friede, T. (2000). Re-calculating the sample size in internal pilot study designs with control of the type I error rate. *Statistics in Medicine*, 19(7):901–911.

Kieser, M., Schneider, B., and Friede, T. (2002). A bootstrap procedure for adaptive selection of the test statistic in flexible two-stage designs. *Biometrical Journal*, 44(5):641–652.

Kimani, P. K., Stallard, N., and Hutton, J. L. (2009). Dose selection in seamless phase II/III clinical trials based on efficacy and safety. *Statistics in Medicine*, 28(6):917–936.

Kirby, S. and Chuang-Stein, C. (2010). Determining sample size for classical designs. In *Handbook of Adaptive Designs in Pharmaceutical and Clinical Development*, pages 6–1–6–25–. CRC Press.

Koenig, F., Brannath, W., Bretz, F., and Posch, M. (2008). Adaptive Dunnett tests for treatment selection. *Statistics in Medicine*, 27(10):1612–1625.

Kropotov, I. D. (2009). *Quantitative EEG, event-related potentials and neurotherapy*. Elsevier/Academic, Amsterdam; Boston; London.

Kudo, A. (1963). A multivariate analogue of the one-sided test. *Biometrika*, 15:403–418.

Lan, K. K. G. and DeMets, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika*, 70:659–663.

Lan, K. K. G. and Wittes, J. T. (2012). Some thoughts on sample size : A Bayesian-frequentist hybrid approach. *Clinical Trials*, 9(5):561–569.

Lancaster, G. A., Dodd, S., and Williamson, P. R. (2004). Design and analysis of pilot studies: Recommendations for good practice. *Journal of Evaluation in Clinical Practice*, 10(2):307–312.

Lang, T., Auterith, A., and Bauer, P. (2000). Trendtests with adaptive scoring. *Biometrical Journal*, 42(8):1007–1020.

Läuter, J. (1996). Exact t and F tests for analyzing studies with multiple endpoints. *Biometrics*, 52:964–970.

Läuter, J., Glimm, E., and Kropf, S. (1996). New multivariate tests for data with an inherent structure. *Biometrical Journal*, 38:1–23.

Läuter, J., Horn, F., Rosolowski, M., and Glimm, E. (2009). High-dimensional data analysis: Selection of variables, data compression and graphics - Application to gene expression. *Biometrical Journal*, 51(2):235–251.

Lehmacher, W. and Wassmer, G. (1999). Adaptive sample size calculations in group sequential trials. *Biometrics*, 55(4):1286–1290.

Lehmacher, W., Wassmer, G., and Reitmeir, P. (1991). Procedures for two-sample comparisons with multiple endpoints controlling the experimentwise error rate. *Biometrics*, 47(2):511–521.

Leroux, B. G., Mancl, L. A., and DeRouen, T. A. (2005). Group sequential testing in dental clinical trials with longitudinal data on multiple outcome variables. *Statistical Methods In Medical Research*, 14(6):591–602.

Lindquist, M. A. (2008). The statistical analysis of fMRI data. *Statistical Science*, 23(4):439–464.

Liptak, T. (1958). On the combination of independent tests. *Akademia Matematikai Kutato Intezetenek Kozlemenyei*, 3:171–197.

Liu, Q., Proschan, M. A., and Pledger, G. W. (2002). A unified theory of two-stage adaptive designs. *Journal of the American Statistical Association*, 97(460):1034–1041.

Logan, B. R. and Tamhane, A. C. (2004). On O'Brien's OLS and GLS tests for multiple endpoints. *Lecture Notes-Monograph Series*, 47:76–88.

Lopes da Silva, F. (2005). EEG analysis: theory and practice. In Niedermeyer, E. and Lopes da Silva, F., editors, *Electroencephalography, 5th edition*. Lippincott Williams and Wilkins.

Lystad, R. and Pollard, H. (2009). Functional neuroimaging: A brief overview and feasibility for use in chiropractic research. *Journal of the Canadian Chiropractic Association*, 53(1):59–72.

Mahalanobis, P. C. (1930). On tests and measures of group divergence. *Journal and Proceedings of the Asiatic Society of Bengal*, 26:541–588.

Mangoubi, R., Desai, M., and Sammak, P. (2008). Non-Gaussian methods in biomedical imaging. In *Applied Imagery Pattern Recognition Workshop, 2008. AIPR '08. 37th IEEE*, pages 1–6.

Marcus, R., Eric, P., and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis*. Academic Press, London; New York.

Matthews, D. E. and Farewell, V. T. (2007). *Using and understanding medical statistics*. Karger, Basel; New York.

Matthews, P. M., Rabiner, I., and Gunn, R. (2011). Non-invasive imaging in experimental medicine for drug development. *Current Opinion in Pharmacology*, 11(5):501–507.

Mehta, C. R. and Patel, N. R. (2006). Adaptive, group sequential and decision theoretic approaches to sample size determination. *Statistics in Medicine*, 25(19):3250–3269.

Mehta, C. R. and Pocock, S. J. (2011). Adaptive increase in sample size when interim results are promising: A practical guide with examples. *Statistics in Medicine*, 30(28):3267–3284.

Miller, F. (2010). Adaptive dose-finding: Proof of concept with type I error control. *Biometrical Journal*, 52(5):577–589.

Minas, G., Aston, J. A., and Stallard, N. (2013). Adaptive multivariate global testing. Technical report, CRISM.

Minas, G., Rigat, F., Nichols, T. E., Aston, J. A., and Stallard, N. (2012). A hybrid procedure for detecting global treatment effects in multivariate clinical trials: theory and applications to fMRI studies. *Statistics in Medicine*, 31(3):253–268.

Mitsis, G. D., Iannetti, G. D., Smart, T. S., Tracey, I., and Wise, R. G. (2007). Regions of interest analysis in pharmacological fMRI: How do the definition criteria influence the inferred result? *Neuroimage*, 40:121–132.

Mulert, C. and Lemieux, L. (2010). *EEG-fMRI physiological basis, technique, and applications*. Springer, Heidelberg; New York.

Müller, H.-H. and Schäfer, H. (2001). Adaptive group sequential designs for clinical trials: Combining the advantages of adaptive and of classical group sequential approaches. *Biometrics*, 57(3):pp. 886–891.

Müller, H.-H. and Schäffer, H. (2004). A general statistical principle for changing a design any time during the course of a trial. *Statistics in Medicine*, 23(16):2497–2508.

Neuhäuser, M. (2001). An adaptive location-scale test. *Biometrical Journal*, 43(7):809–819.

O'Brien, P. C. (1984). Procedures for comparing samples with multiple endpoints. *Biometrics*, 40:1079–1087.

O'Brien, P. C. and Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, 35(3):549–556.

Ogawa, S., Tank, D., Menon, R., Ellermann, J., Kim, S., Merkle, H., and Ugurbil, K. (1992). Intrinsic signal changes accompanying sensory stimulation: Functional brain mapping with magnetic resonance imaging. *Proceedings of the National Academy of Sciences USA*, 89(13):5951–5955.

O'Hagan, A. and Stevens, J. W. (2001). Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making*, 21.

Peacock, J. and Peacock, P. J. (2011). *Oxford handbook of medical statistics*. Oxford University Press, Oxford.

Perlman, M. (1969). One-sided testing problems in multivariate analysis. *Annals of Mathematical Statistics*, 40:549–567.

Phillips, M. L. (2012). Neuroimaging in psychiatry: Bringing neuroscience into clinical practice. *The British Journal of Psychiatry*, 201(1):1–3.

Pihlajamäki, M. and Sperling, R. A. (2008). fMRI: Use in early Alzheimer's disease and in clinical trials. *Future Neurology*, 3(4):409–421.

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199.

Pocock, S. J. (1997). Clinical trials with multiple outcomes: A statistical perspective on their design, analysis, and interpretation. *Controlled Clinical Trials*, 18(6):530–545.

Pocock, S. J., Geller, N. L., and Tsiatis, A. A. (1987). The analysis of multiple end-points in clinical-trials. *Biometrics*, 43(3):487–498.

Poldrack, R. A., Mumford, J. A., and Nichols, T. E. (2011). *Handbook of functional MRI data analysis*. Cambridge University Press, Cambridge.

Pong, A. and Chow, S.-C. (2010). Overview of adaptive design methods in clinical trials. In *Handbook of Adaptive Designs in Pharmaceutical and Clinical Development*, pages 1–1–1–19–. CRC Press.

Posch, M. and Bauer, P. (1999). Adaptive two stage designs and the conditional error function. *Biometrical Journal*, 41(6):689–696.

Posch, M., Bauer, P., and Brannath, W. (2003). Issues in designing flexible trials. *Statistics in Medicine*, 22(6):953–969.

Proschan, M. A. (2009). Sample size re-estimation in clinical trials. *Biometrical Journal*, 51(2):348–357.

Proschan, M. A. and Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics*, 51(4):1315–1324.

Raichle, M. E. (2000). 2 - A brief history of human functional brain mapping. In Toga, A. W. and Mazziotta, J. C., editors, *Brain Mapping: The Systems*, pages 33–75. Academic Press, San Diego.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional data analysis*. Springer, New York.

Rao, K. Ramamohan, K. D. N. H. J. J. (2010). *Fast Fourier transform algorithms and applications*. Springer, Dordrecht; New York.

Rencher, A. C. (2002). *Methods of multivariate analysis*. J. Wiley, New York.

Rencher, A. C. and Christensen, W. F. (2012). *Methods of multivariate analysis*. J. Wiley, New York.

Ripley, B. D. (1981). *Spatial statistics*. Wiley, New York.

Sanei, S. and Chambers, J. (2007). *EEG signal processing*. John Wiley & Sons.

Sankoh, A. J., Huque, M. F., and Dubey, S. D. (1997). Some comments on frequently used multiple endpoint adjustment methods in clinical trials. *Statistics in Medicine*, 16(22):2529–2542.

Schmidli, H., Bretz, F., and Racine-Poon, A. (2007). Bayesian predictive power for interim adaptation in seamless phase II/III trials where the endpoint is survival up to some specified timepoint. *Statistics in Medicine*, 26(27):4925–4938.

Schwarz, A. J., Becerra, L., Upadhyay, J., Anderson, J., Baumgartner, R., Coimbra, A., Evelhoch, J., Hargreaves, R., Robertson, B., Iyengar, S., Tauscher, J., Bleakman, D., and Borsook, D. (2011a). A procedural framework for good imaging practice in pharmacological fMRI studies applied to drug development #1 : Processes and requirements. *Drug Discovery Today*, 16(1314):583–593.

Schwarz, A. J., Becerra, L., Upadhyay, J., Anderson, J., Baumgartner, R., Coimbra, A., Evelhoch, J., Hargreaves, R., Robertson, B., Iyengar, S., Tauscher, J., Bleakman, D., and Borsook, D. (2011b). A procedural framework for good imaging practice in pharmacological fMRI studies applied to drug development #2: Protocol optimization and best practices. *Drug Discovery Today*, 16(15-16):671 – 682.

Shih, W. J., Ohman-Strickland, P. A., and Lin, Y. (2004). Analysis of pilot and early phase studies with small sample sizes. *Statistics in Medicine*, 23(12):1827–1842.

Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika*, 73(3):751–754.

Smith, A. M., Longo, C. A., Fried, P. A., Hogan, M. J., and Cameron, I. (2010). Effects of marijuana on visuospatial working memory: An fMRI study in young adults. *Psychopharmacology*, 210(3):429–438.

Spiegelhalter, D. J. and Freedman, L. S. (1986). A predictive approach to selecting the size of a clinical trial. *Statistics in Medicine*, 5:1–13.

Spiegelhalter, D. J., Freedman, L. S., and Blackburn, P. (1986). Monitoring clinical trials: conditional or predictive power? *Controlled Clinical Trials*, 7:8–17.

Squire, L., Roberts, J., Spitzer, N., and Zigmond, M. (2008). *Fundamental neuroscience, third edition*. Elsevier / Academic Press, Amsterdam; Boston.

Stallard, N. and Todd, S. (2011). Seamless phase II/III designs. *Statistical Methods in Medical Research*, 20(6):623–634.

Steiger, A. and Kimura, M. (2010). Wake and sleep EEG provide biomarkers in depression. *Journal of Psychiatric Research*, 44(4):242–252.

Stein, C. (1945). A two-sample test for a linear hypothesis whose power is independent of the variance. *The Annals of Mathematical Statistics*, 16(3):243–258.

Stewart, J. L., Coan, J. A., Towers, D. N., and Allen, J. J. B. (2011). Frontal EEG asymmetry during emotional challenge differentiates individuals with and without lifetime major depressive disorder. *Journal of Affective Disorders*, 129(1-3):167–174.

Talairach, J. and Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain : 3-dimensional proportional system : An approach to cerebral imaging*. G. Thieme ; Thieme Medical Publishers, Stuttgart; New York; New York.

Tang, D.-I., Geller, N. L., and Pocock, S. J. (1993). On the design and analysis of randomized clinical trials with multiple endpoints. *Biometrics*, 49(1):23–30.

Tang, D.-I., Gnecco, C., and Geller, N. L. (1989a). An approximate likelihood ratio test for a normal mean vector with nonnegative components with application to clinical trials. *Biometrika*, 76(3):577–583.

Tang, D.-I., Gnecco, C., and Geller, N. L. (1989b). Design of group sequential clinical trials with multiple endpoints. *Journal of the American Statistical Association*, 84(407):776–779.

Thompson, J. L. and Ebersole, J. S. (1999). Long-term inpatient audiovisual scalp EEG monitoring. *Journal of Clinical Neurophysiology*, 16(2):91–99.

Tsiatis, A. A. and Mehta, C. (2003). On the inefficiency of the adaptive design for monitoring clinical trials. *Biometrika*, 90(2):367–378.

Vandemeulebroecke, M., Bretz, F., Pinheiro, J., and Bornkamp, B. (2010). Adaptive dose-ranging studies. In *Handbook of Adaptive Designs in Pharmaceutical and Clinical Development*, pages 11–1–11–16–. CRC Press.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, 16(2):117–186.

Wang, S. K. and Tsiatis, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics*, 43(1):193–199.

Wei, L. J. and Durham, S. (1978). The randomized play-the-winner rule in medical trials. *Journal of the American Statistical Association*, 73(364):840–843.

Westfall, P. H., Krishen, A., and Young, S. S. (1998). Using prior information to allocate significance levels for multiple endpoints. *Statistics in Medicine*, 17(18):2107–2119.

Westfall, P. H. and Young, S. S. (1993). *Resampling-based multiple testing : Examples and methods for P-value adjustment*. Wiley, New York.

Whitcher, B. and Matthews, P. (2006). Noninvasive brain imaging for experimental medicine in drug discovery and development: Promise and pitfalls. *International Journal of Pharmaceutical Medicine*, 20:167–175(9).

Whitehead, J. (1997). *The design and analysis of sequential clinical trials*. J. Wiley & Sons, Chichester; New York.

Wise, R. G. and Tracey, I. (2006). The role of fMRI in drug discovery. *Journal of Magnetic Resonance Imaging*, 23:862–876.

Wittes, J. and Brittain, E. (1990). The role of internal pilot studies in increasing the efficiency of clinical trials. *Statistics in Medicine*, 9(1-2):65–72.

Wong, D. F., Tauscher, J., and Grunder, G. (2008). The role of imaging in proof of concept for CNS drug discovery and development. *Neuropsychopharmacology*, 34(1):187–203.

Worsley, K. J., Evans, A. C., Strother, S. C., and Tyler, J. L. (1991). A linear spatial correlation model, with applications to positron emission tomography. *Journal of the American Statistical Association*, 86(413):55–67.

Zhu, H. J. and Hu, F. F. (2010). Sequential monitoring of response-adaptive randomized clinical trials. *Annals of Statistics*, 38(4):2218–2241.